I know where in Europe you will
want to study next Semester:

Towards Adaptive Retrieval and Recommendation of Higher Education Programmes

Thijs Putman – 0534518
Department of Mathematics and Computer Science
Eindhoven University of Technology

A thesis submitted for the degree of
Master of Science in Business Information Systems

May 2010

Assessment Committee

Dr. M. Pechenizkiy (Supervisor)
Department of Mathematics and Computer Science
Databases & Hypermedia Group

Dr. T. G. K. Calders
Department of Mathematics and Computer Science
Databases & Hypermedia Group

Prof. Dr. P. M. E. De Bra
Department of Mathematics and Computer Science
Databases & Hypermedia Group

Dr. J. J. L. Schepers
Department of Industrial Engineering and Innovation Sciences
Innovation, Technology Entrepreneurship & Marketing Group

Dr. M. Voorhoeve
Department of Mathematics and Computer Science
Architecture of Information Systems Group

Executive Summary

MastersPortal.eu is a website which aims to provide detailed information about all Master's programmes in Europe. It has seen tremendous growth, both in number of visitors and programmes listed, over the past three years. With this growth it has encountered a multitude of challenges, both technical and commercial in their nature.

In order to increase the value MastersPortal adds to both its student visitors and its university customers, an investigation was started into the performance of the current website. From this investigation it became apparent that the bounce-rate of visitors coming from Google, the single biggest referrer for the MastersPortal.eu website, is very high: Many of the visitors referred from Google only view a single page on MastersPortal before they leave.
In an attempt to decrease the bounce-rate, the implementation of a programme recommender on the MastersPortal.eu website is proposed. The reasoning is that if an incoming visitor is presented with a number of programmes relevant to him; he will be less inclined to return to Google. Instead, the visitor continues his quest for information on MastersPortal.
In order to determine the overall effect of implementing a recommender system and to come to the optimal recommendation approach, an online controlled experiment was designed and executed.

From this online experiment, it became clear that implementing a recommender system has a substantial, reducing, effect on the bounce-rate. As a result, an increase in visitor retention of over 90% is achieved by the best performing recommender approach.
Furthermore, the recommender approach chosen has a significant effect on the performance of the recommender system. From the first experiment it is concluded that both a content-based and a collaborative recommender approach perform well on the MastersPortal.eu website. Based on this conclusion a content-based recommender has been deployed on the MastersPortal.eu website since March of this year.

In order to better understand the effectiveness of the different recommender approaches, the results of the first experiment are further analysed by looking at contextual factors. These factors provide insight into the situation of each visitor. From the analysis executed it follows that classifying visitors based upon contextual factors often provides substantially different results when compared to the overall, uncategorised, results. The most important contextual factor taken into account during this thesis is the geographic origin of the visitor.
A second experiment was executed to verify the findings with respect to the visitor's geographic origin. The influence of the visitor's geographic origin detected during the first experiment is strongly verified by the second experiment. Additional effects were also observed. These seem to be caused by the high pace of change within both the MastersPortal.eu website and its environment.

A potentially valuable improvement to the recommender system on the MastersPortal.eu website would be to enable it to adapt itself based upon the contextual factors identified during this thesis. The results from the second experiment are used to explore this possibility. The hypothetical adaptive system analysed, stabilises in line with expectations: It allows some groups of visitor to see a recommender which for them performs better, but overall would not get preference.

Although no functional adaptive recommender system is constructed as part of this thesis, the results of the exploratory analysis offer a concrete direction for future developments within MastersPortal. A substantial amount of additional work is required before any kind of context-aware recommender system can be implemented though.

In conclusion, placing a programme recommender on the MastersPortal.eu website has a substantial reducing effect on the bounce-rate of visitors coming in from Google. Both a content-based and a collaborative approach provide similar performance in case of the MastersPortal.eu website. Further potential lies in finding the optimal, potentially adaptive, combination of multiple recommender approaches based upon a visitor's contextual factors.
As a result of the work performed as part of this thesis, the MastersPortal.eu website now has a significantly lower bounce-rate; through this increased retention of visitors MastersPortal has achieved a 14% growth in revenue potential. Moreover, all visitors now have access to a useful recommender system that helps to optimise their study choice.

Acknowledgements

Special thanks to my supervisor Dr. Mykola Pechenizkiy for his guidance and suggestions throughout this project and for allowing me the freedom to pursue this thesis such that it could be combined with my responsibilities at MastersPortal.
Many thanks to team over at MastersPortal, Edwin and Magnus in particular, for allowing me to complete this thesis amidst the organised chaos that ensued while building out our hobby project into a “real” company.

Thanks to all members of Studievereniging Interactie, and its European counterpart ESTIEM, who I have had the pleasure to work, and party, with over the past years. They have moved my focus across the Dutch borders into Europe and have laid the foundation for what has become the MastersPortal.eu website.

Thanks to Bas Bemelmans together with whom I have done much coursework during my time at university. We have kept each other in line and focussed and Bas has been an invaluable sounding board throughout this project.
Finally, thanks to all my room-mates at the `s Gravesandestraat for their moral support, especially during the last few long weeks of writing this thesis.

Eindhoven, May 11th 2010

Table of Contents

1. Introduction

This chapter provides a general introduction to this thesis. It includes background on the MastersPortal.eu website and an introduction into the circumstances of the problem faced. The chapter furthermore includes a summary of the results of this thesis.

1.1 About MastersPortal.eu

MastersPortal.eu aims to construct an online database in which students from all over the World can search European Master's programmes. It was started at the end of the 2006 as a student initiative in response to the Bologna Process.
As part of the Bologna Process, 47 European countries aim to unify their individual systems of higher education. The signatories of the Bologna Process strive to promote student mobility, increase the attractiveness of the European higher education as a whole and improve synergies with the Anglo-Saxon system of higher education.

An important effect of the Bologna Process is a vastly increased focus on international student mobility, both in between the signatory countries and from nations outside of Europe. Furthermore, the traditional integrated programmes taught in continental Europe are split up in two phases: The Bachelor's and Master's phase. This creates a whole new decision moment for students at the end of their Bachelor's phase, as well as tens of thousands of new Master's programmes that need to be brought to the attention of students from all over the world.

MastersPortal was started based on the realisation that, though international student mobility is an important part of the Bologna Process, there was not a single unified resource available for students striving to study a Master's programme somewhere in Europe. As such, the MastersPortal.eu website was started with its main goal to promote studying a Master's degree in Europe. MastersPortal aims to reach as many students as possible, world-wide.

The first version of the MastersPortal.eu website went live in June of 2007. At that point its database contained 1.500 Master's programmes and received just over 5.000 monthly visits. Two years later, in May of 2009, the MastersPortal database contained 11.000 Master's programmes. In that same month, the website received 255.000 visits.
Nearing the end of this thesis, in March of 2010, the MastersPortal.eu website received nearly $1.000.000$ monthly visits and contained over 15.000 programmes. The total number of Master's programmes in Europe relevant to the MastersPortal is estimated to be around 40.000.

The student initiative from 2006 has since forth been transformed into StudyPortals B.V., a limited liability company employing nine full-time and six part-time employees. StudyPortals is currently expanding into related markets. New portals, such as the Bachelors-, PhD- and ScholarshipPortal are being developed. Furthermore, the MastersPortal.eu website itself is expanding its scope to include information in multiple European languages.
In order to maintain a clear focus, this thesis limits itself to English-language Master's programmes presented on the MastersPortal.eu website. Information in other languages and the Bachelors- and PhDportal, both launched in January of 2010, are disregarded.

The revenue model for the MastersPortal.eu website is based on selling web-based advertising campaigns and providing sponsored search-results. The revenue model thus focuses on maximising the exposure of university customers. For MastersPortal it is thus vital to increase both the number of visitors it receives and the number of pages viewed by each of these visitors. A steady increase in both leads to a healthy and sustainable growth in revenue received through the products currently offered.
In the future additional services, such match-making between qualified students and interested universities, will also become available. This will shift the focus of the revenue model to be more referral based.

The single most important challenge for the MastersPortal.eu website is to unlock the wealth of information contained within in its ever growing database. Over the past three years, much emphasis has been placed on the back-end systems, providing for efficient addition and management of Master's programmes. The website front-end, used by the student visitors, has not been properly (re)designed since its initial inception in June of 2007.
As such, the growing number of programmes makes it increasingly difficult for students to find the most relevant information. Students who are not able to find relevant information are often put off and leave the MastersPortal.eu website before their information need is fulfilled. Other initiatives providing similar information are only a few clicks away.

This thesis thus starts with the realisation that the MastersPortal.eu website requires better *Information Retrieval* technologies for it to become the *de facto* information resource on European higher education programmes. To provide a better picture of what changes are required, an investigation is started into the current performance of the MastersPortal.eu website; looking for potential areas of improvement.

1.2 Visitor Statistics

As noted in the introduction, both the MastersPortal.eu website and its database have seen major growth over the past three years. In this chapter I will provide an overview of this growth in more detail and as such provide the numerical grounding for the problem description found in the next chapter of this introduction.
The statistics discussed in this chapter range from October 2008 up to and including October 2009, at which point work started on the experiments presented later on in this thesis. The statistics presented in this chapter are generated by the third-party application Webalizer[1]; they are based upon the web server log-files of the MastersPortal.eu website.

The metric used to measure the number of visitors to the MastersPortal.eu website is *visits*. For the current statistical analysis, a visit is defined as a set of requests from a single IP address, separated by at least sixty minutes from the previous or next set of requests from the same address. Only requests for actual pages are included; requests for images and other resource files are not.
All non-human visitors have, for as far as possible, been excluded from the statistics. Due to the stateless nature of the HTTP protocol it is difficult to properly exclude *all* non-human visitors from the statistics presented.

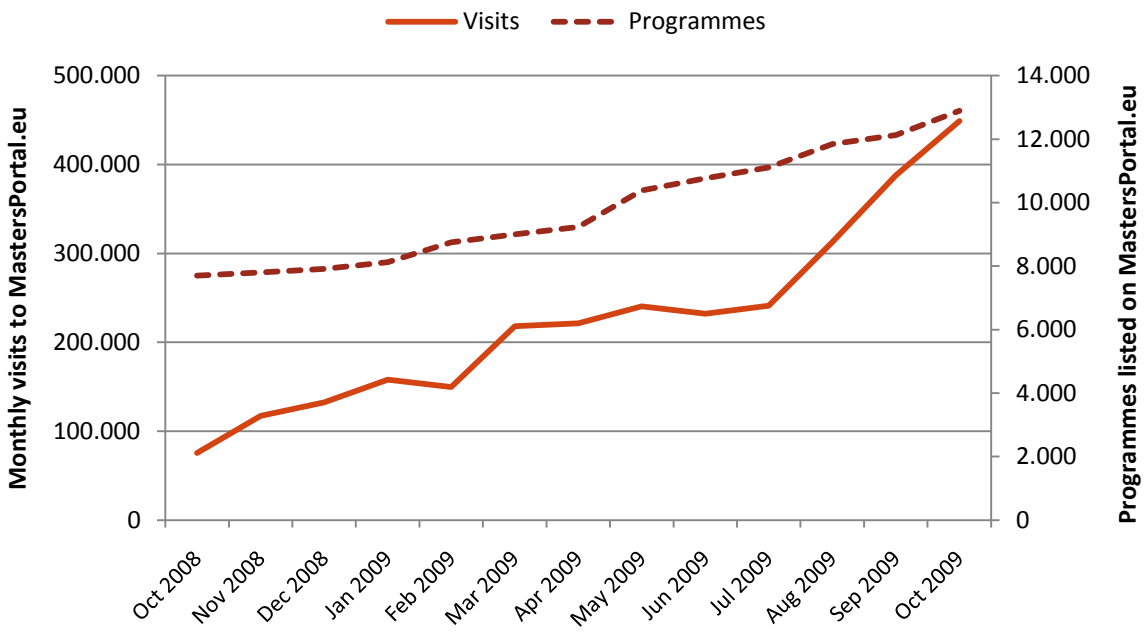


Figure 1: Monthly visits to MastersPortal.eu and the number of Master's programmes listed.

The graph in *Figure 1* maps the increase in number of visits against the increase in Master's programmes listed on the MastersPortal.eu website. Initially we see a strong correlation between the number of Master's programmes and the amount of visits.
Around July of 2009, the number of visits jumps ahead of the number of programmes listed. This trend has remained apparent ever since. At the time of writing, the number of Master's programmes listed has stabilised at around 15.000, while the number of visits stills shows a strong month-over-month increase, leading up to nearly one million visits in March of 2010.

[1] http://www.stonesteps.ca/projects/webalizer/

Initially, the growth of the MastersPortal.eu website was caused, for a large part, by an increase in its information offering. Month-over-month visitor growth follows the increase in programmes closely, as can be seen in *Figure 1*. For the last months of 2009, the increase in visits has become less related to the number of programmes offered.
It is important to note this development, as it goes to show MastersPortal is becoming less dependent upon programme growth to increase its visitor numbers. This is an important factor in its commercial relevance. It indicates an increase in value delivered to the university customers. With an equal number of programmes listed, more visitors mean more exposure for the universities.

1.2.1 The Impact of Google

Ever since the launch of the MastersPortal.eu website, Google has played an important part in its growth. When we take the graph from *Figure 1* and add the number of visits originating from Google's search-engine we see this very clearly, as presented in *Figure 2*.

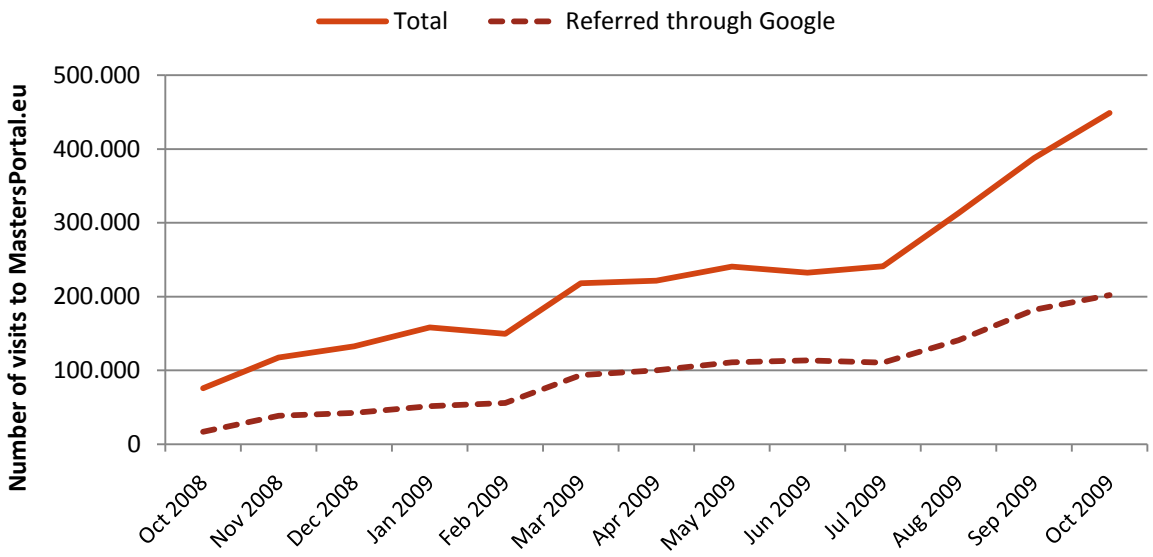


Figure 2: Monthly visits to MastersPortal.eu; total and those incoming through Google's search-engine.

From July of 2009 onward, visitors from Google's search-engine account for around 45% of the monthly visits to the MastersPortal.eu website. There are several other major search-engines linking to the MastersPortal.eu website, such Yahoo! and Bing, but none of them have an impact anywhere near as big as that of Google. Yahoo! and Bing combined, for example, provide less than one percent of the total monthly visits.
For the MastersPortal.eu website, search-engine traffic thus effectively equals Google traffic. For the remainder of this document, whenever referring to search-engine traffic, it may be implicitly assumed this is traffic coming in from Google's search-engine.

A short, qualitative, analysis of the visitors referred to the MastersPortal.eu website through Google shows a very diverse set of queries, both generic and very specific. The most used search-terms refer to either the "MastersPortal" itself, or to the more general concept of "studying in Europe".
Apart from these generic search-terms, an extremely varied set of more specific terms is present, leading users from Google to each and every one of the Master's programmes listed on the MastersPortal.eu website.

1.2.2 Visit Duration

As noted in the introduction, apart from a steady increase in number of visitors, an important commercial metric for the MastersPortal.eu website is the amount of pages viewed per visit. It is an indication of how relevant visitors of the MastersPortal.eu website find the information. Apart from providing an indication of relevance, the visit duration metric is also important to the commercial side of the MastersPortal. Much of the revenue generated is linked directly to the number of pages viewed by MastersPortal.eu's visitors.

A substantial month-over-month increase in visitor numbers is already being achieved, but for MastersPortal.eu to become sustainable in the long term, it is equally important to increase its visit duration. The average visit duration per month over the period running from October of 2008 until October of 2009 is graphed in *Figure 3*.

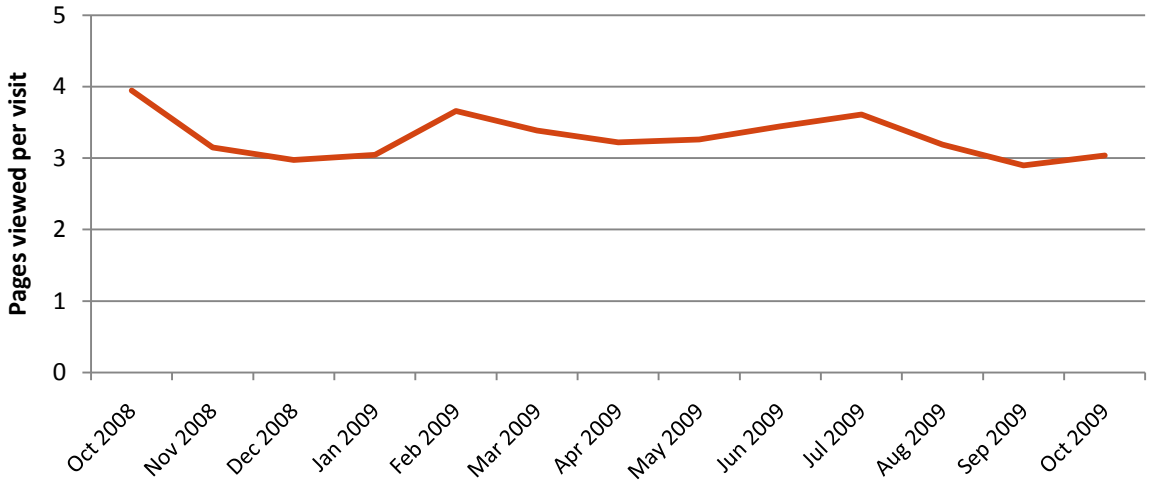


Figure 3: Number of pages viewed per visit on MastersPortal.eu.

The visit duration in *Figure 3* shows a less positive picture. Although the number of visits is rising; outpacing the programme growth, the number of pages viewed per visit is not increasing at all. One could actually argue it is slowly declining.

This stable visit duration points towards the conclusion that MastersPortal is unable to provide visitors with a satisfying experience. Visitors referred by external websites do not stay on the MastersPortal.eu website. Improving this metric is an important part in increasing the commercial performance MastersPortal.eu and ensuring its sustainability.

1.2.3 Bounce- and Revisit-Rate

To gain a further insight into why the visit duration at MastersPortal.eu is not increasing, statics covering the last four weeks of the previously analysed period, ranging from the 5th of October up to and including the 1st of November of 2009, are analysed in more detail.

For this detailed analysis, two metrics are taken into account. Firstly the bounce-rate metric, which is based upon the number of visitors referred to the MastersPortal.eu website who only view a single page. After having viewed this single page, these visitors return to the referring website and do not return for at least sixty minutes, the duration of a visit. These visitors are considered to have "bounced". A high bounce-rate, just as low visit duration, points to the fact that visitors do not find the information they are looking for.

Secondly the revisit-rate metric, which is defined as the number of visitors that, after an initial visit to MastersPortal.eu, return at a later point in time (c.q. at least sixty minutes later) for a second visit. A low revisit-rate again indicates lack of relevance; it furthermore goes to show that visitors need to be convinced during their initial visit. Both the bounce-rate and the revisit-rate are thus strongly related to the relevance of the information presented.

As mentioned earlier, Google is the single largest referrer for the MastersPortal.eu website. As such, visitors referred by Google are the main subject of the bounce-rate and revisit-rate investigations. Using visitors from only a single external entity helps to reduce variance brought into the analysis. As Google dwarfs all other referrers in size by a significant margin, it is the logical choice.

To provide a frame of reference concerning the bounce-rate, a second bounce-rate analysis is performed. This analysis calculates the bounce-rate for the combination of all other external referrers that are *not* search-engine. Leaving the other search-engines out of the analysis provides a good indication of how referrals from search-engines, Google in this case, perform when compared to other external referrals.

The results of the analysis are presented in *Appendix A*. From the analysis executed in this appendix, *Figure 4* below is constructed.

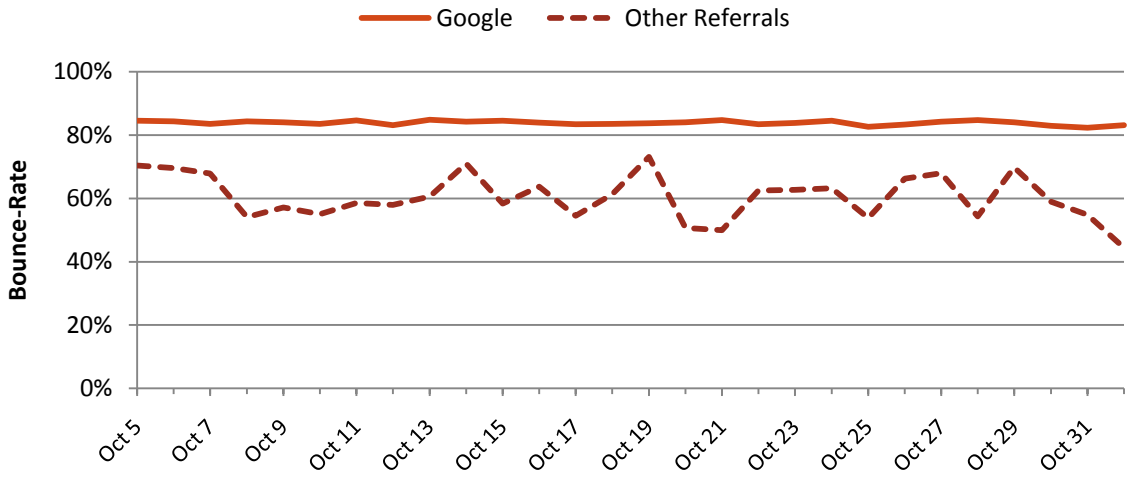


Figure 4: Bounce-rate of Google visitors and visitors from other, non search-engine, externals referrers.

The figure above provides an overview of the bounce-rate for visitors coming from Google's search-engine compared to visitors coming from other referrals. The comparison in *Figure 4* clearly shows the bounce-rate for visitors coming in from Google is structurally much higher than that of the other external referrals.
The bounce-rate for Google visitors in the above graph is steady around 84% ($\mu = 83{,}88\%$; $\sigma = 0{,}70\%$). The bounce-rate for the other external referrals is a little less stable around 65% ($\mu = 64{,}47\%$; $\sigma = 7{,}26\%$), but it is clearly well below the Google bounce-rate. The instability is caused by the much smaller dataset available when computing the bounce-rate for the other referrals.

Less than 20% of visitors coming from Google actually stay on the MastersPortal.eu website and view more than a single page. This indicates most of the visitors coming from Google either think they cannot find, or really do not find, what they were looking for.
Assuming both types of referral visitors come to the MastersPortal.eu website because they are interested in "studying in Europe", the bounce-rate for visitors referred by Google seems high. When compared to referrals from other external websites, the bounce-rate for Google referrals is consistently nearly 20 percent-points higher.
Decreasing the bounce-rate for visitors referred by Google has a major effect on the overall bounce-rate of the MastersPortal.eu website *and* its commercial performance. Even though visitors referred by Google might not be similar to those referred by other websites, we do think the large difference in performance between the two groups points to the potential for a significant reduction in the Google bounce-rate.

Finally, in *Appendix A2*, some details are presented concerning the revisit-rate of visitors coming from Google. Looking at the numbers in the appendix we see that once bounced, only 18% of the Google referrals come back for a second visit.
This underlines the fact that apart from a relatively high bounce-rate, Google visitors also have a low revisit-rate. It is thus of great importance to convert visitors referred by Google during their initial visit, or more specifically, during their initial page view. For many of the visitors we will not get a second chance.

1.3 Problem Description and Proposed Solution

From the discussion of MastersPortal.eu's visitor statistics in the previous section, a clear problem emerges. In this section, the problem uncovered is summarised and the solution we aim to develop for this problem is introduced.

1.3.1 Problem Description

Many visitors of the MastersPortal.eu website do not come directly to its homepage; they do not use the MastersPortal search-engine while looking for Master's programmes. Instead, these visitors are referred by Google directly to a page containing detailed information on only one of the many Master's programmes contained within the MastersPortal database.

For many of these visitors the large amount of additional information available to them through the MastersPortal.eu website is not apparent. Over 80% of these visitors are lost immediately after viewing their initial page. These visitors only view a single Master's programme before returning to Google.

For nearly all of these visitors, additional pages with relevant information exist within the MastersPortal database. But, instead of utilising the information in MastersPortal's database, these visitors continue their quest on Google. Furthermore, less than 20% of these visitors return to the MastersPortal.eu website for a second visit at a later point in time.

> **MastersPortal.eu** does not succeed in convincing a large portion of its first-time visitors to stay on its website after their initial page view. The message that the MastersPortal.eu website has much more to offer than the initial page viewed is not conveyed effectively. As a result of this much potential is lost, both commercially for MastersPortal and intrinsically for the student visitor who does not find the information he is looking for.

1.3.2 Proposed Solution

To solve MastersPortal's problem a solution needs to be devised which offers incentive to visitors to stay on the MastersPortal.eu website, even though they did not find exactly what they were looking for on the initial page of their visit.

On many e-commerce websites a common way to increase revenue is the prominent inclusion of recommendations for related products. Instead of navigating away, the visitor's attention is drawn to these related products. Visitors do not return to their external referrer and continue their shopping there; they stay on the e-commerce website and offer it yet another opportunity to sell its merchandise.

Providing a recommendation for related Master's programmes seems a feasible solution to the problem of the MastersPortal.eu website. By using the initial Master's programme viewed by a visitor as an implicit indication of this visitor's information need, a recommender system can provide the visitor with additional relevant information.

By implementing a recommender system, the MastersPortal.eu website will be better able to convince its visitors of its added value. Through showing a set of related Master's programmes to a visitor, this visitor is made aware of the fact that the MastersPortal database contains additional relevant information. This in turn might prevent the visitor from "bouncing" back to his external referrer.

1.4 Methodology and Results

This section provides an overview of the methodological approach to the research conducted during this thesis as well as a summary of the results achieved.
Denning, as cited in (Nunamaker, Chen, & Purdin, 1990), states that engineers generally agree that progress is achieved primarily through posing problems and systematically following a process to construct systems that solve these problems. Based on this engineering principle the authors propose the *systems development* methodology for information systems research. As such, systems development is used as a basis for the methodological approach to the research conducted during this thesis.

The systems development research process as proposed by (Nunamaker, Chen, & Purdin, 1990) consists of five stages: Construct a conceptual framework; develop a system's architecture; analyse and design the system; build the system; observe and evaluate the system.
For this thesis, the process is condensed into three stages by combining the first three steps defined by the authors. These steps are combined into a single stage considering relevant background knowledge. Our research methodology thus consists of three stages: gather background knowledge; construct a prototype system; evaluate the prototype system.
During the first stage, the current state of affairs in literature concerning recommender systems is analysed. Based on this analysis, recommender approaches are selected, goals are defined and an experimental study is designed. During the second stage a prototype recommender system is constructed based on the background knowledge gathered. During the third and final stage of a research cycle the prototype system is evaluated using both real-world tests and a simulation study. Focus is placed on the final two steps of the research process. Our aim is to provide a practical solution the problem at hand and to evaluate its effectiveness. The diagram in *Figure 5* provides an overview of the use of the *systems development* research process within this thesis.

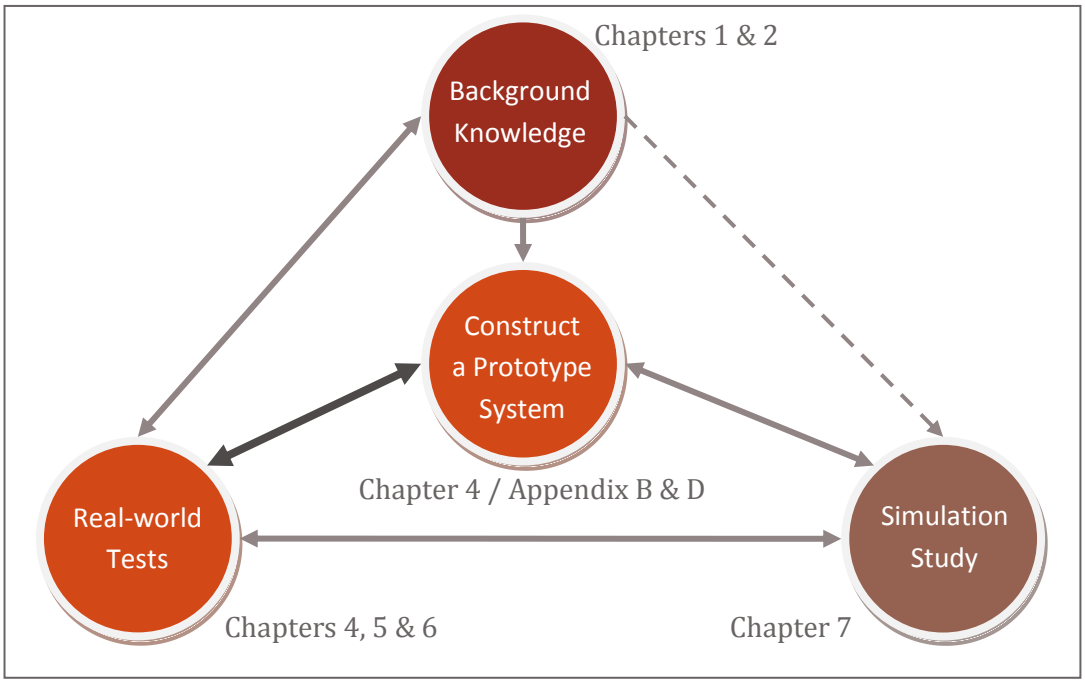


Figure 5: The thesis' systems development research process adapted from (Nunamaker, Chen, & Purdin, 1990)

The following sections shortly summarise the results presented in the remaining chapters of this thesis. An exhaustive discussion of the experiments and analyses executed is provided in *Chapters 4* through *7*; a detailed discussion of the results and their implications is provided in *Chapter 8*.

Chapter 4: Recommender Experiment

To come to the best possible recommender approach for the MastersPortal.eu website, several alternatives are implemented and evaluated using an online controlled experiment. The results of this first experiment show that providing visitors of the MastersPortal.eu website with a programme recommendation has a distinctly positive effect on their bounce-rate. A drop in bounce-rate, and as a result a rise in visitor retention, of over 90% is achieved by both the content-based and collaborative recommender approaches.

The nature of the recommender approaches implemented as part of the recommender system appears to play an important role in the bounce-rate reduction. When visitors are provided with a random set of programme recommendations their bounce-rates remain unchanged; providing a relevant recommendation, on the other hand, decreases the bounce-rates significantly.

Chapter 5: Contextual Factors

Intuitively, visitors of the MastersPortal.eu website should not be considered a single homogenous group. MastersPortal's visitors come for all parts of the world and as such from many different walks of life. Therefore, the data gathered during the first experiment is analysed in an effort to unearth differences in performance between these various groups of visitors.

Visitors are classified based upon several contextual factors. During the further analysis of the results of the first experiment, three contextual factors are taken into account: The geographic origin of the visitor; the query the visitor entered on Google before being referred to MastersPortal; the academic discipline of the Master's programme viewed by the visitor.

A fourth contextual factor, the screen resolution of the visitor, is introduced but deliberately omitted from the first analysis on contextual factors. This fourth factor is taken into account in the later discussion on the feasibility of an adaptive recommender system.

During the analysis of the contextual factors several contextual visitor categories are found that perform in total opposition to the overall result of the first experiment.

Chapter 6: Contextual Factors Experiment

To verify the conclusions of the first experiment and to confirm the influence of the contextual factors, a second experiment is set up and executed. To limit complexity of the second experiment, it aims to verify the effects of a single contextual factor: The geographic origin of a visitor. The second experiment verifies both the overall result of the first experiment and the existence of differences within the geographic origin contextual factor. The experiment furthermore provides evidence of additional influences of the geographic origin of a visitor that were not visible before.

A proper understanding of the preferences of different contextual visitor categories matters greatly if we want to further optimise the performance of the recommender system on the MastersPortal.eu website. It appears to be difficult to predict visitor preferences in advance or to provide generalisations based on the contextual effects observed.

Chapter 7: Feasibility of an Adaptive Recommender System

Based on the results of both experiments an additional analysis, concerned with the feasibility of constructing an adaptive recommender system utilising the contextual factors identified in this thesis, is undertaken. The opportunity for this analysis became apparent based on the results of the experiments executed (it was not originally intended to be a part of this thesis). The analysis aims to satisfy, at least to some extent, the curiosity raised by the results of the experiments and to provide concrete directions for future developments within MastersPortal.

The analysis shows that the data gathered during the second experiment provides good grounds for a stable adaptive system. The adaptations proposed by the hypothetical system are furthermore in line with the expectations raised by the results of both experiments.
A system which utilises a single, static, recommender approach often provides certain groups of visitors with a recommendation that does not perform most favourable for them. An adaptive recommender system on the other hand is able to optimise its recommendations based upon the contextual categories of each visitor.
The additional analysis indicates there is the potential to successfully implement an adaptive recommender system on the MastersPortal.eu website: The contextual factors discussed in this thesis provide good grounds for adaptation and clear performance differences between different contextual visitor categories are observed. Much additional research is though required before any kind of adaptive system can be implemented on MastersPortal.

2. Background and Design Decisions

Before starting a detailed discussion on the experiment's results, this chapter provides background on key concepts used throughout this thesis. It is by no means an exhaustive exploration of all concepts, but it serves to introduce the main concepts and explain important choices made throughout this thesis.

2.1 MastersPortal Database

This first part of the background discussion focuses on the structure of the MastersPortal database. It provides an introduction into how information within the MastersPortal database is structured and how elements are related. A simplified overview of the database is provided in *Figure 6*.

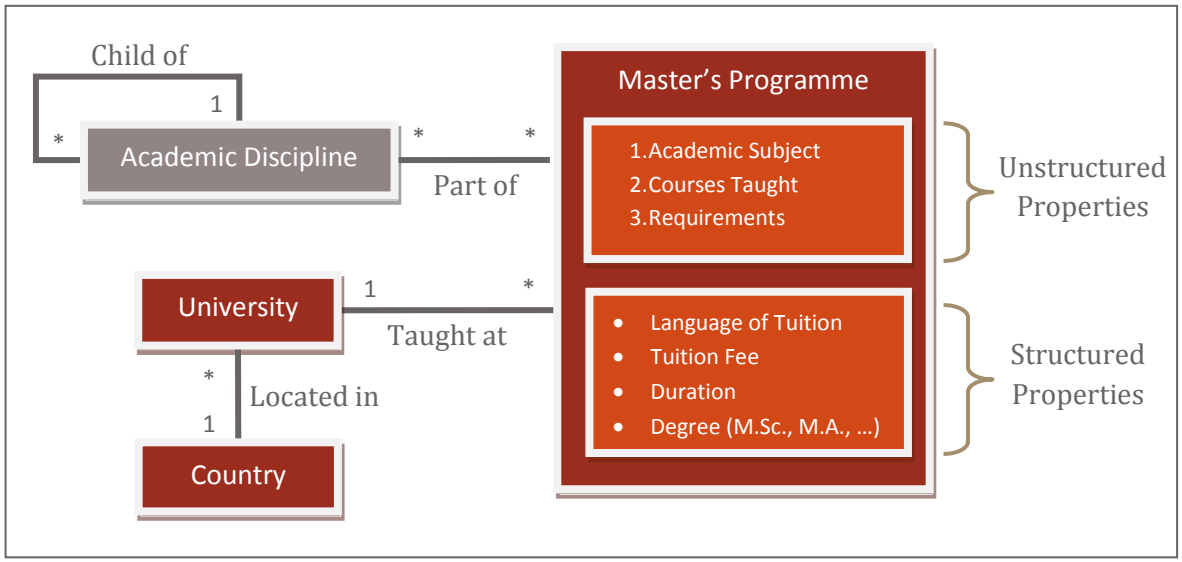


Figure 6: Simplified overview of the MastersPortal database

The overview provided in *Figure 6* is incomplete. It leaves out many additional elements and relationships found in the MastersPortal database and it omits many of the properties found on each of these elements. The goal of the simplified overview is to provide a clear picture on how the concepts relevant to this thesis are related.

The MastersPortal database consists of a collection of Master's programmes that are contained within both a geographic and an academic (c.q. contextual) hierarchy.
The geographic hierarchy consists of the university the programme is taught at and the country in which this university is located. This hierarchy is used internally to manage the Master's programmes and the administrative access rights assigned to each programme.
The academic hierarchy serves to categorise Master's programmes based upon their academic content. Each programme has a unique position in the geographic hierarchy, but it can be part of multiple academic disciplines, that themselves can be part of other academic disciplines. The disciplines thus form a multi-level hierarchy.

Each Master's programme contains a set of properties, some of which are structured and some of which are unstructured. Concerning the academic contents of each Master's programme the unstructured properties are of the greatest importance; they pertain to the subject, courses taught and requirements of the Master's programme.
The structured properties of a Master's programme deal more with its prerequisites. These properties provide little information on the academic contents of the programme; they serve to indicate the "limiting factors" of each programme in a structured way. Elements such as the programme's language of tuition and its tuition fees are not directly related to academic content. They are simply filters to exclude results inherently out of one's reach.
Similarly, the geographic hierarchy can be viewed as a prerequisite. Many students are limited in a geographical sense and only want to see programmes from regions where they can, or want to, study. Within these regions the students still need to resort to the unstructured properties to judge the academic contents of a programme. For example, a student might want to study in Scandinavia because of its low tuition fees. The student is thus limited to the northern European part of the geographic hierarchy. After applying this initial restriction the student still needs to consider the unstructured properties of all programmes in Northern Europe to find those that best matches his academic interests.

From experience within the MastersPortal we know that students looking for a Master's degree place high importance in the academic contents of the programme. What makes a Master's programme a good choice for a student, is that it connects to their prior education and covers a topic of interest to them. Within this decision there are of course limitations on location, duration and cost, but the academic contents of the programme are central.
Although this observation holds true for Master's programme, we have noticed this is not always the case. For students considering a Bachelor's programme for example the academic content is of less importance. Many students, fresh from high school, are more interested in going to a city nearby, going where their friends do or learning something very "general". For future initiatives of MastersPortal the situation as described above might thus be different.

While setting up the MastersPortal database, much time was invested into the construction of an exhaustive and mutually exclusive list of academic disciplines. The goal of this list is to provide an inherent contextual structure to all current and future programmes included in the database. The current set of disciplines within the MastersPortal database is provided in *Appendix K*. All Master's programmes in the database are categorised into at least one of these disciplines by their administrator. This administrator, often an employee of the university where the programme is taught, can be seen as a domain expert on the programme's content.
The current list of academic disciplines is not free from problems though. Administrators often find it difficult to properly categorise their programmes. Differences in opinion exist on the exact nature of certain disciplines and not all disciplines are as mutually exclusive as hop. The level of distinction offered by a classification based on the academic disciplines is limited.

The programme information most valuable to the MastersPortal's visitors is contained within the programme's unstructured properties. If we are to implement a recommender system on the MastersPortal.eu website we thus cannot solely rely on the hierarchy, geographic or academic, present in the database. Both of the hierarchies simply do not take into account the information most valued by the visitors.

In the next section of this background discussion we will thus look at automated and, consequently, more advanced recommender approaches.

Although the academic hierarchy in itself is not the most useful source of a programme recommendation, it can certainly help improve recommender performance in the future. Many sources in literature indicate the categorisations made through this kind of hierarchy can be used to enhance the performance of recommender systems:
"[We] claim that human knowledge is crucial for the effectiveness of text categorisation and text retrieval" (Yang & Chute, 1994). "Our experiments clearly indicate that the categorisation process is effective. It improves the retrieval performance compared with no categorisation. It also achieves the retrieval performance equivalent to the results using manual categorisation." (Lam, Ruiz, & Srinivasan, 1999).
The academic interests of most students visiting the MastersPortal.eu website are limited to a handful of the disciplines present in the academic hierarchy. Even though the disciplines are quite broad, by only considering a couple, the number of potentially relevant programmes is greatly reduced. By providing a recommendation within this reduced set of programmes, the most relevant programmes from an already more relevant subset are retrieved.
Future improvements to a recommender system on the MastersPortal.eu website can thus benefit from the inherent academic hierarchy present in its database. This is a feature that, for the remainder of this thesis, will not be taken into account.

2.2 Recommender Systems

The proposed solution to the problem of the MastersPortal.eu website is the implementation of a recommender system. In this chapter we explore the current state of affairs with respect to online recommender systems in an e-commerce setting.

Recommender systems have been in use on the Web for over 15 years. In 1999 many of the then top e-commerce website employed a variety of recommender systems to try and increase their revenue (Schafer, Konstan, & Riedl, 1999). The authors define three goals for recommender systems in an e-commerce environment: They should convert browsing visitors into buyers, lead to cross-selling and improve customer loyalty. As the MastersPortal.eu website does not sell directly to its visitors, the latter two goals are most relevant to our situation. In a world where alternative websites are only a few clicks away, improving visitor loyalty is especially important.
The use of recommender system in an e-commerce setting is a much researched and established field. Within the market the MastersPortal.eu website operates in this is less the case. The other major players[2] in the field have yet to invest in recommender technologies. Although recently community features have gained traction, many of MastersPortal's competitors have never invested in anything beyond the most basic information retrieval technologies. This is thus a field in which the MastersPortal.eu website can further differentiate itself.

Looking at the types of recommender systems proposed and implemented over the years, there is roughly speaking a distinction to be made between two types: Content-based and collaborative recommender systems (Balabanovic & Shoham, 1997).
According to (Balabanovic & Shoham, 1997) content-based systems recommend "[text documents] based on a comparison between their content and a user profile. Data structures for both of these are created using features extracted from the text of the documents. Often some weighting scheme is used which gives high weights to discriminating words".
The collaborative approach to recommendation is very different: "Rather than recommend items because they are similar to items a user has liked in the past, we recommend items other similar users have liked".
In recent years recommender performance has been further improved by combining both content-based and collaborative approaches into hybrid recommender system (Adomavicius, Sankaranarayanan, Sen, & Tuzhilin, 2005). In the example provided by these authors this is done by "learning and maintaining user profiles based on content analysis" and subsequently "comparing the resulting profiles to determine similar users in order to make collaborative recommendations".

Taking into account the above definitions and the goal of this thesis, it seems best to investigate both a content-based recommender and collaborative recommender approach. Based on the results of both approaches we can decide how to proceed; most likely by working towards a more hybrid approach in the future. It does not seem sensible to immediately implement a hybrid recommender. Our primary goal is to see *whether* implementing a recommender system has any effect and *what* this effect is. By implementing a hybrid recommender right away we make it difficult to judge which of the two approaches works best on the MastersPortal.eu website and why this is the case.

[2] Relevant other initiatives, similar in size and target audience, are "Find a Masters" (http://www.findamasters.com/), Educations.com (http://www.educations.com) and GradSchools.com (http://www.gradschools.com).

In the next three sections the baseline recommendation and the content-based and collaborative recommenders to be introduced on the MastersPortalo.eu website are discussed.

2.2.1 Baseline Recommendation

Before discussing the content-based and collaborative recommender approaches, we use the discussion of the MastersPortal database, from *Chapter 2.1*, to define a baseline recommendation. The baseline recommendation is based upon the academic hierarchy present in the MastersPortal database. This is something we know visitors of the MastersPortal.eu website are interested in. The baseline recommendation can thus be used to judge the added value of the more advanced recommender approaches discussed in the next sections.

When a visitor views a Master's programme, the baseline recommendation provides a list of the Master's programmes that have the largest overlap in academic disciplines as compared to the reference programme.

$$Sim(D_R, D_P) = \frac{\sum_i (D_{R,i} * D_{P,i})}{|\{d \in D_R : d = 1\}|} \quad \text{with} \quad D = (d_1, d_2, \ldots, d_n) \quad \text{and} \quad d \in \{0,1\}$$

This method takes the *n-tuple* of disciplines attached to a reference programme (D_R) and computes the overlap with the *n-tuple* of disciplines attached to another programme (D_P). Note that the *n-tuple* of disciplines has the same length for each programme.
The baseline recommendation essentially functions as a form of "query by example". The academic disciplines can be used to filter the interests of a visitor. If a visitor would use MastersPortal's search-engine he might apply this filter manually by selecting the disciplines of his interest. The baseline recommender simply attempts to automate this process.

For the purpose of applying the above method, the academic hierarchy is flattened. Each programme is categorised into all parent discipline(s) of its current discipline(s). Apart from simplifying the procedure, this modification also helps to better weight the similarity score in cases where only a partial match, a match at a higher level in the hierarchy, is found.
A recommendation is constructed by comparing the reference programme against all other programmes in the MastersPortal database. The *top-n* most relevant programmes based on $Sim(D_R, D_P)$ are used as a recommendation for the reference programme.

2.2.2 Content-Based Recommender

Nearly all information on the academic contents of Master's programmes in the MastersPortal database is available in the form of unstructured text. As such, the content-based recommender is concerned with feature extraction from these unstructured texts.
The content-based recommender system takes a single Master's programme as a reference. It applies a weighting scheme to formalise the unstructured information on its academic contents and creates a relevancy ranking between the reference programme and the other Master's programmes found in the database.

The content-based recommender thus requires a weighting scheme which is able to create a formal representation of unstructured text. First and foremost amongst the possible schemes is the tf-idf approach: "The [tf-idf] vector space approach and the cosine similarity function have been applied to several text classification applications and despite the algorithm's unquestionable simplicity, it performs competitively with more complex algorithms" (Pazzani & Billsus, 2007).

During a previous study[3] executed within a MastersPortal setting, several weighting schemes were examined. The study compares the tf-idf approach against the BM-25 approach (Robertson, Walker, Hancock-Beaulieu, Gull, & Lau, 1995) and pivoted normalisation (Singhal, Buckley, & Mitra, 1996) with as goal to optimise the retrieval of Master's programmes from a body of unstructured and largely irrelevant information.
The conclusion of the study is that both BM-25 and pivoted normalisation perform slightly worse than a tf-idf implementation in the retrieval of relevant documents for MastersPortal. Their computational performance is nonetheless much better than that of the tf-idf approach. For our current purpose, constructing a recommender system, the computational advantages are not considered to outweigh a greater retrieval performance.
The above result combined with the fact that the tf-idf approach has been previously implemented within MastersPortal leads to its selection as basis for the content-based recommender. In the future MastersPortal should certainly investigate additional content-based retrieval techniques. The quick review done as part of this thesis indicates several feasible alternatives to the tf-idf approach exist. A further investigation might unearth a recommender approach that outperforms the tf-idf approach on all facets relevant to MastersPortal's situation.

The tf-idf weighting method takes the term vector of a reference programme (V_R) and compares it to another programme's term vector (V_P). A recommendation is constructed by comparing the reference vector against all other vectors in the MastersPortal database. The *top-n* most relevant programmes based on $Sim(V_R, V_P)$ are used as a recommendation for the reference programme.

$$Sim(V_R, V_P) = \frac{\sum_{i=1}^{P}(w_{R,i} * w_{P,i})}{\sqrt{\sum_{i=1}^{N} w^2{}_{R,i}} * \sqrt{\sum_{i=1}^{N} w^2{}_{P,i}}} \quad \text{with} \quad w = tf \cdot ln(df) \qquad \text{(Garcia, 2006)}$$

Before constructing the reference and programme vectors, all English-language stop words are removed and stemming according to Porter's algorithm (Porter, 2006) is applied. As this thesis focuses solely on English-language Master's programmes, no facilities are implemented to accommodate other languages. Full details on the implementation of the tf-idf content-based recommender are provided in *Appendix D1*.

[3] The results of this study are provided in *Appendix L*.

2.2.3 Collaborative Recommender

By viewing a set of Master's programmes on the MastersPortal.eu website, visitors classify this group of Master's programmes as related to their information need. "Collaborative filtering systems are built on the assumption that a good way to find interesting content is to find other people who have similar interest, and then recommend titles that those similar users like" (Breese, Heckerman, & Kadie, 1998). The groupings created by one visitor might thus be of interest to other visitors with a similar information need.
Collaborative recommender systems are widely deployed in online scenarios and it has been empirically shown that they provide good recommendations: "Collaborative recommendation has been demonstrated empirically, and has been widely adopted commercially" (O'Mahony, Hurley, Kushmerick, & Silvestre, 2004). The authors do note that "there is no general theoretical explanation of the conditions under which a particular collaborative recommendation application will succeed or fail". It is thus interesting to see how the collaborative recommender approach performs within the confines of the MastersPortal.eu website.

A collaborative recommender functions by "[retrieving] customers who have expressed similar preferences to the target, and then recommend items that were liked by the retrieved customers" (O'Mahony, Hurley, Kushmerick, & Silvestre, 2004). By combining previous behaviour of visitors of the MastersPortal.eu website, many groups of related programmes can be constructed. Using these groupings a relevancy ranking can be made for each Master's programme in the database. This ranking can be used as the basis for a collaborative recommender system.
To gather the groupings, usage patterns are harvested from web server log-files. "While click-through data is typically noisy and clicks are not `perfect' relevance judgments, the clicks are likely to convey some information" (Joachims, 2002). The author goes on to note that "[the] key insight is that such click-through data can provide training data in the form of relative preferences".

The collaborative method applied during this thesis takes all groupings in which both a reference programme (r) and another programme (p) are present. A recommendation is constructed by taking the fraction of the groupings in which both r and p are present compared to the number of groupings in which either is present. The *top-n* most relevant programmes based on $Sim(r,p)$ are used as a recommendation for the reference programme.

$$Sim(r,p) = \frac{|O|}{|R \cup P|} \qquad \text{with} \qquad O = \{G \in All : p \in G; r \in G\} \qquad \text{and}$$

$$P = \{G \in All : p \in G\}, \quad R = \{G \in All : r \in G\}, \quad All = \{All\ programme\ groupings\}$$

A large amount of historical visitor data is required to construct a set of groupings sufficiently large to use as the basis for the collaborative recommender system. (Joachims, 2002) Notes that using web server log-files to gather this data is not an ideal solution. Although for this thesis it is the only feasible solution: Within MastersPortal, information on visitor behaviour has never been stored in a formalised way. There is no database of previous visitor actions or preferences. Harvesting click-through information from the raw web-server log-files is thus all that remains. Practical details on the actual implementation of the collaborative recommender are available in *Appendix D2*.

An added issue with the implementation of the collaborative recommender is that it is open to various external biases. For example, visitors using MastersPortal's search-engine are biased by its

presentation and ordering of results. Once implemented, the collaborative recommender itself becomes a source of bias.

The collaborative recommender system implemented as part of this thesis should thus be viewed as relevant for the investigations performed during this thesis. Afterwards, if the collaborative approach performs well, a new implementation needs to be devised.
A potential future approach lies in a system along the lines of Facebook's "Like" button[4]. Through this button visitors indicate they "like" a specific webpage. It is essentially a relevance feedback system, whose data can be further used to construct better groupings of Master's programmes. As MastersPortal intends to implement community-based features in the near future, this is a possibility to take into consideration.
Additionally, statistical modelling can be used to infer additional groupings from a limited amount of information on visitor behaviour (Breese, Heckerman, & Kadie, 1998).

[4] http://developers.facebook.com/docs/reference/plugins/like/

2.3 Context and Adaptation

Within an online scenario it is possible to provide a customised experience to each visitor. As noted by Jeff Bezos, CEO of Amazon: “If I have 2 million customers on the Web, I should have 2 million stores on the Web” (Schafer, Konstan, & Riedl, 1999).
If we want to apply these kinds of personalisation on the MastersPortal.eu website, an adaptive recommender system is required. As defined by (Balabanovic, 1997) an adaptive recommender system “seeks to adapt to its users, providing increasingly personalised recommendations over time”. An adaptive recommender thus automatically modifies itself based on the interests of a visitor in an attempt to better serve him.

According to Balabanovic, many adaptive systems are constructed by combining content-based and collaborative recommender systems. Constructing an adaptive recommender is thus one of the many ways of implementing a hybrid recommender system. As discussed previously, working towards the implementation of a hybrid recommender system is a potential future direction for MastersPortal.
The results of the experiments executed during this thesis indicate adaptive features can improve the performance of MastersPortal’s recommender system. Both the content-based and collaborative recommenders have their own merits and when properly combined perform better than the sum of their parts. Therefore it is interesting to study if an adaptive recommender system is feasible within the setting of the MastersPortal.eu website.
Based on the above realisation an additional analysis is added to this thesis. This analysis investigates the potential of implementing an adaptive recommender system on the MastersPortal.eu website. This part of the thesis should be considered an addition to its main topic; not an integral part of it. The discussion on context and adaptation provided in this section is as a result less elaborate than the further background provided in this chapter.

The main goal of the recommender system on the MastersPortal.eu website is to reduce the bounce-rate of visitors referred by search-engines, Google in particular. Our adaptive system thus needs to adapt based on the information available during the first page view of a visit.
Adaptations thus need to be based on the profile constructed for each visitor based on the implicit information these visitors provide at the start of their first visit to the MastersPortal.eu website. This implicit information can be described as a visitor’s context.
According to (Dey, 2001), “context is any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves”.
As we have seen in the introduction to this thesis, many first-time visitors to the MastersPortal.eu website bounce back to Google after viewing a single page. If the first experience of a visitor more personalised, and thus more relevant to this visitor, we might be able to further prevent him from bouncing.

To further personalise the experience offered by MastersPortal’s recommender system we can incorporate the detection of contextual factors into the system and adapt the recommendations based on these factors. The ensuing recommender system becomes *context-aware*, it “uses context to provide relevant information and/or services to the user, where relevancy depends on the user’s task" (Dey, 2001).
Dey defines three possible uses context-aware systems: “presentation of information and services to a user”, “automatic execution of a service for a user” and “tagging of context to information to

support later retrieval". Within this thesis we will focus mostly on the first usage, the presentation of information to best suit the specific situation of a visitor.

(Domingues, Jorge, & Soares, 2009) Note that "existing contextual recommender systems typically use contextual information as a label for segmenting/filtering sessions, using them to build the recommendation model" Although the authors themselves introduce a different approach to contextual recommenders, for the purpose of this thesis we will simply use contextual information to classify visitors. By classifying visitors based on this information and subsequently adapting the recommendation based upon these categorisations, our recommender system becomes context-aware.

Both (Domingues, Jorge, & Soares, 2009) and (Adomavicius, Sankaranarayanan, Sen, & Tuzhilin, 2005) underline the point that contextual information matters in improving recommender performance. Adomavicius literally states "context matters". Both authors furthermore state that using multiple contextual factors produces better results. The authors do note that using contextual information does not always improve performance. We will thus need to remain critical on the effectiveness of the use of contextual information.
(Domingues, Jorge, & Soares, 2009) Further stress that not all contexts are relevant all of the time: "not every contextual dimension significantly affects a given recommendation task". We will thus also need to ensure that the contexts we select are helpful in MastersPortal's specific situation.

Based on the definition of context provided by (Dey, 2001), the contextual factors selected as part of this thesis concern themselves with the "situation" of a visitor to the MastersPortal.eu website. This situation can be related to the visitor's interests, both in general or specific to higher education programmes. It can be related to the personal profile of the visitor, what is his background and where is he from. Technical details such as his browser type and screen resolution can be taken into account as well. Many other more complicated contexts are also possible. For example, by looking at social networking sites we can attempt to determine the profile of the visitor based upon his relationships to other, already known, visitors.

The possibilities are virtually limitless, which according to (Domingues, Jorge, & Soares, 2009) makes "identifying rich contextual dimensions not an easy task". We will need to be critical in our selection and evaluation of potential contexts. The effects of each context should be considered very carefully and based upon this consideration a decision should be made on whether or not the contextual factor adds enough value to be included in an adaptive recommender system on the MastersPortal.eu website.

2.4 E-commerce Performance Metrics

In order to properly evaluate the results of implementing different recommender approaches, an appropriate performance metric needs to be chosen.
The standard approach to evaluating an information retrieval system is to consider the number of relevant and irrelevant documents retrieved by the system (Manning, Raghavan, & Schütze, 2009). A relevance judgement is made by a user based upon his information need. This relevance judgement is based on the number of documents retrieved that fulfil the information need.

When the aim of an information retrieval system is to return all relevant, and no irrelevant, documents from a collection, its performance is often judged through the *precision* and *recall* measures. Precision is the fraction of retrieved documents that are relevant; recall is the fraction of relevant documents that are retrieved (Manning, Raghavan, & Schütze, 2009).
For many prominent web applications precision and recall are not very effective. Both online search-engines and online recommender systems return a set of ranked documents. Visitors cannot be expected to make a full analysis of the entire ranked set before they judge its relevance. The visitor's judgement is thus based upon the top k highest ranked documents in the set. To cope with this fact, (Manning, Raghavan, & Schütze, 2009) consider the concept of "precision at k" which allows one to measure the relevance of the top k ranked results.
Although precision at k is defined with an online environment in mind, it remains difficult to measure outside of a laboratory environment. If we ask visitors for an explicit relevance judgement the mere presence of the question might influence their answer. Furthermore, many additional biases are introduced by directly asking visitors for their feedback. Finally, the precision at k value is not all that relevant during this thesis. It tells us nothing about the effect the recommender system has on the revenue generated by MastersPortal.

The problem faced by the MastersPortal.eu website is similar to that found on many e-commerce websites. As such we turned our attention to e-commerce literature to find a more suitable performance metric.
For recommender systems used on e-commerce websites the value of a recommendation can easily be measured in terms of the revenue it generates. Bad recommendations will not entice customers to buy. As the MastersPortal.eu website does not sell products to its student visitors, it is difficult to use metrics based upon these kinds of "return-on-investment" judgements.
For the MastersPortal.eu website revenue is not directly generated by the student population visiting it; income is generated through university partnerships. An increase in exposure for the university partners leads to an increase in revenue. As such MastersPortal aims to increase the number of page views generated by its student visitors.

A different metric is required for use on the MastersPortal.eu website. In literature two alternative metrics are found: Firstly the *click-through-rate*, which is based on the number of "physical" clicks attracted by a recommendation. The second alternative is the *bounce-rate* which is an indirect metric aiming to provide better insight into the final relevance judgement a visitor makes concerning the content he is linked to.

2.4.1 Click-through-Rate

The click-through-rate metric stems from the online advertisement market. Most banner advertisements are primarily sold and evaluated on a "per-click" basis. Within the search-engines of Google, Microsoft and Yahoo! this metric is an important steering factor for banner placements (Richardson, Dominowska, & Ragno, 2007).

The click-through-rate (CTR) is defined as the number of clicks a link receives dived by the number of times the link is viewed. The click-through rate of a link can be easily and reliably measured. It forms a clear metric which, up to some extent, is able to capture the relevance of a link: Only visitor who think the link will benefit them click on it.

$$CTR = \text{number of } clicks/number\ of\ impressions$$

Within the search-engines mentioned above, the click-through-rate is used directly as a revenue-based metric. Search-engines are interested in providing relevant advertisements, leading up to sales for their customers. Their revenue model on the other hand is purely based upon the number of clicks generated by the advertisements of these customers. As such, the click-through-rate is the preferred performance metric for the search-engines.
According to (Sculley, Malkin, Basu, & Bayardo, 2009) "One limitation [of the click-through-rate] is that it fails to capture the user's evaluation of the ensuing experience on the landing page because the landing page is not visible prior to the click". This is an important drawback for the MastersPortal.eu website. Our aim is to increase the number of pages viewed. This is done best by providing visitors with a satisfying experience throughout their visit; especially after they click on a link which is presented to them as a relevant alternative to the current page.

The click-through-rate is thus not the best metric for use on the MastersPortal.eu website during this thesis. It is strongly related to the previously discussed revenue-based metrics and offers no good ultimate relevancy judgement. In order to increase the number of pages viewed, measuring clicks on links alone is not enough: the MastersPortal.eu website needs to "measure" if visitors are satisfied with the full information offering they receive.

2.4.2 Bounce-Rate

An alternative to the click-through-rate metric is the bounce-rate of visitors. The bounce-rate is a relatively unstudied metric; it has only gained attention in the past few years. This is in part caused by the fact that it is more difficult to measure than the click-through-rate, especially in the banner advertisement scenarios mentioned in the previous section.
The bounce-rate is closely related to the click-through-rate metric. In a recent study done at Google several million advertisements shown in response to search queries were compared. Based on this comparison it was concluded that "advertisements with very low observed bounce-rate have very high [click-through-rate]" (Sculley, Malkin, Basu, & Bayardo, 2009). According to the authors a very strong inverse correlation between the bounce-rate and the click-through-rate exists.

The bounce-rate (BR) of visitors from external referrers is defined as the number of successful visits divided by the total number of visits from this referrer.

$$BR = successful\ visits\ /\ total\ visits$$

The definition of a *successful visit* depends highly on the setting of the website on which the metric is used. An e-commerce website might consider a visit successful once a purchase is made; a social networking website might do the same if a new user registers. In general, each website has a set of desirable actions it wants its visitors to undertake. A visit is considered successful if one or more of these actions is completed by the visitor.

In order to gather an accurate bounce-rate metric, search-engines need to share information with their banner advertisement customers, or they need to develop accurate estimators for the bounce-rate based upon other metrics: "search engine provider[s] can observe a user's behaviour on the search engine itself, but cannot make observations after the user has clicked through to an advertisement" (Sculley, Malkin, Basu, & Bayardo, 2009).
As opposed to merely measuring the interest of a visitor in the link presented, the bounce-rate measures the impact of the landing page on the user's behaviour. According to (Sculley, Malkin, Basu, & Bayardo, 2009) "high bounce-rates may indicate that users are dissatisfied with page content or layout or that the page is not well aligned to their original query".

According to this definition, the bounce-rate can be considered an enhanced version of the click-through-rate metric, largely negating the objections posed to its use on the MastersPortal.eu website in the previous section.
Apart from a programme recommendation, the MastersPortal.eu website offers two other important elements on most of its landing pages: Visitors can search the entire MastersPortal database and they can get in touch with a university contact person. Both of these actions provide further exposure to universities and thus generate revenue. Whether a visitor executes any of these additional actions can be influenced by the recommendation provided. Measuring only the click-through-rate of the recommendation does not capture this effect; measuring the bounce-rate of visitor does.
The bounce-rate metric is closely related to MastersPortal's commercial goal of increasing the number of pages viewed by its visitors. A visitor who does not bounce will, by definition, generate more page views than a visitor that does bounce. The bounce-rate is thus more appropriate than the click-through-rate if we aim to measure an increase in the number of pages viewed by visitors.

(Sculley, Malkin, Basu, & Bayardo, 2009) Do note that: "any practical method of observing user bounces is prone to some error". This is something we need to take into account while interpreting the results of our experiments. The case of the MastersPortal.eu website is less complicated than the situations referred to by these authors. We attempt to measure the bounce-rate of visitors to a single website, over which we have a large degree of control. We can thus perform a more precise measurement than is possible in the search-engine scenario investigated by (Sculley, Malkin, Basu, & Bayardo, 2009).
The results based on the bounce-rate metric are though still slightly open to interpretation. The potential errors introduced through this fact are of far less significance than the problems with the click-through-rate metric identified in the previous section.

2.5 Online Controlled Experiments

According to (Kohavi, Longbotham, Sommerfield, & Henne, 2008), the Web "provides an unprecedented opportunity to evaluate ideas quickly using controlled experiments". As opposed to offline experimentation, in an online experiment we have the possibility to quickly test features on actual visitors and gather results of their actual behaviour.
Many initiatives, such as Microsoft's Experimentation Platform[5] and Google's Website Optimizer[6], are leading the way in commoditising online controlled experiments. At Amazon "data trumps intuition", indicating the world's largest online retailer attaches great value to the results of their online controlled experiments (Kohavi, Longbotham, Sommerfield, & Henne, 2008).

Many of the largest players in the Web thus use online controlled experiments as part of their day-to-day business and decision-making. Many of the dilemmas faced by Microsoft, Google and Amazon also play within MastersPortal, albeit on a reduced scale. As such, controlled experiments can also benefit MastersPortal.
The decision to use online controlled experiments as part of this thesis should thus not only be considered within the confines of this thesis: We also hope to introduce controlled experiments as an accepted decision-making technique within MastersPortal, creating an "experimentation culture".

The most commonly used type of online controlled experiment is a univariate experiment, referred to as "A/B testing" in the (online) advertisement industry (Manning, Raghavan, & Schütze, 2009). The goal of such an experiment is to decide between two implementations of a single feature. Univariate experiments can be extended to evaluate multiple alternatives of a single feature. The overall procedure remains similar in these cases, though the analyses executed differ somewhat.
Another method of experimentation gaining popularity in an online setting is multivariate testing. These kinds of experiments consider changing multiple features during a single experiment. These setups are unnecessarily complicated within the scope of this thesis and as such are not further considered.

2.5.1 Hypothesis Testing

To properly evaluate the results of an online controlled experiment, some statistical grounding is required. In this section we will shortly discuss the procedure applied to analyse the results of the online controlled experiments executed as part of this thesis.

As noted, the most common online testing approach consists of univariate experiment in which the current situation, the control, is compared to a potential improvement, the treatment. A decision is made by measuring a performance metric for both situations and applying a t-test to determine which of the two situations performs most favourably.
Within the setting of this thesis, multiple control groups need to be compared against multiple treatment groups. Although this has no major ramifications on the design of the experiment, it does influence its evaluation. When drawing conclusions on pair-wise relations within more than two groups it is not advisable to use a standard t-test (McClave, Benson, & Sinicich, 2001).
The standard t-test disregards potential interactions between the elements and increases the chance of *Type I error*, the possibility of rejecting the null hypothesis when it is actually true. As such, a multiple comparison procedure needs to be employed (Cooper & Schindler, 2003).

[5] http://exp-platform.com/
[6] http://www.google.com/websiteoptimizer/

A multiple comparison procedure is used to construct a ranking between control and treatment groups. After applying the procedure, the differences between the group means may be compared. The procedure consists of two steps.
Firstly, a set of range tests is executed whose purpose is to find homogenous groups within the dataset. The second step is the actual multiple comparison procedure which provides an indication of the actual difference in means between the groups identified.
The multiple comparison procedure is a type of ANOVA test and as such several assumptions are required: All measurements need to be randomly selected from a normal population; all groups involved in the analysis should have equal variances; all measurements within a group should be independent of each other.

The bounce-rate performance metric discussed in the previous section is binomially distributed. As the ANOVA test requires normally distributed measurements, the bounce-rate thus needs to be approximated with a normal distribution. According to (Montgomery & Runger, 2003) this is allowed by using the following approximation for the average and standard deviation:

$$\mu = np, \quad \sigma = np\,(1 - p) \qquad \text{with} \qquad \mu > 5, \quad \sigma > 5$$

In these equations n is the number of measurements and p the success probability, c.q. the number of participants who did *not* bounce. The approximation does not work well for small *n*. Conversely, the larger the number of participants, the better the approximation fits.
For all analyses executed as part of this thesis both μ and σ remain well within the limits of the approximation. The smallest groups of participants considered as part the experiments contain around 1.000 individuals; with an infeasible high bounce-rate of 99% both μ and σ still remain well clear of the threshold values.

Applying a normal approximation does lead to a second issue: The normally distributed measurements do not have equal variances. To overcome this limitation, Bonferroni's method is used while applying the multiple comparison procedures. Bonferroni's method is one of the most stringent approaches to multiple comparisons; it is a conservative choice. This allows for pair-wise comparison of groups with unequal averages and unequal variances. (Cooper & Schindler, 2003)
We expect a large amount of data to be available to evaluate the results of the experiments; as such a conservative approach can be used without harming the significance of the conclusions.

In *Chapters 6* and *7* regular t-tests are applied as in these chapters only two groups are compared. The normal approximation described above is used in these cases too. As a result of this approximation, the basic Student t-test cannot be applied; it again assumes equal variances. To counteract this Welch's approach to t-testing is used. This is an adaption of Student's t-test which allows it to cope with differences in variance between the two normal distributions compared (Welch, 1947).

2.5.2 Validation

As with all experimentation, the results of online controlled experiments need to be validated. Most textbooks on experiment design distinguish between internal and external validity.
According to (Cooper & Schindler, 2003) problems of internal validity "can be solved by the careful design of experiments". External validity on the other hand is "largely a matter of generalisation, which, in a logical sense, is an indicative process of extrapolating beyond the data collected". The diagram in *Figure 7* provides an overview of the relationship between external and internal validity.

As a matter of definition it is not possible to design an experimental study which provides both kinds of validity in full. By using an online controlled experiment, we are though able to reach an optimal position in the middle of the validity range displayed in *Figure 7*. Online A/B testing allows us to reach the maximum of both internal and external validity.
Just as with a laboratory experiment, we have a large degree of control over most of the variable factors that occur during the experiment. This leads to a high internal validity. By deploying our experiment in a "live" environment we are able to test the hypotheses against real visitors. These visitors are not aware that they are participating in an experiment and as such behave naturally. This leads to a high external validity.

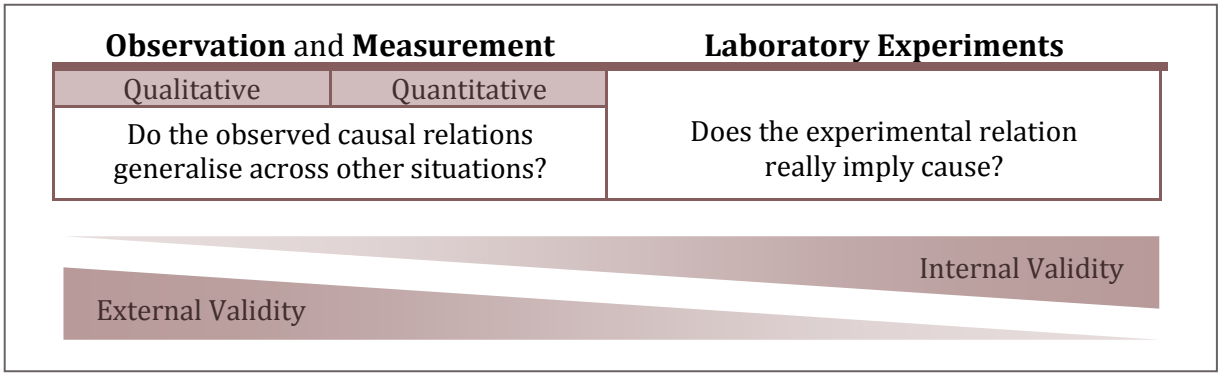


Figure 7: External versus internal validity (Cooper & Schindler, 2003) (McKirnan, 2010)

Apart from not telling the visitors that they are participating in an online experiment there is not much we can do to further increase external validity. In order to further secure internal validity we need to carefully analyse the results of the experiment.
Several components of internal validity are discussed by (Cooper & Schindler, 2003). The most relevant factors concerning the experiments executed as part of this thesis are the "history" and "instrumentation" components.
The history component of validity ensures unidentified factors do not influence the result of the experiment. Checking this component effectively ensures the experiment is controlled. Throughout this thesis, history validity is maintained by rigorously filtering both non-human participants and asserting that human participants do not violate the controlled conditions of the experiment. Further details on this filtering approach are discussed in *Chapter 4.3.2*. The validity of the filtering process itself is considered as part of *Appendix F1*.
Instrumentation is concerned with the validity of the data gathering instruments used as part of an experiment. Instrumentation validity can be assured by comparing the results of the experiment against results from another data source, preferably generated through a separate process. If both sets of data correlate as expected, and as such confirm each other's results, validity of the experiment's instrumentation can be assumed.

According to (Crook, Frasca, Kohavi, & Longbotham, 2009), “multiple [live] A/A tests must be run in order to have confidence whether biased robots exist in the data”. Within the field of online experimentation the effects of non-human visitors form a major threat to validity. A further component of internal validity is thus concerned with preventing this influence. This component is closely related to the “history” component defined by (Cooper & Schindler, 2003). It does have a specific cause and potential set of solutions, which is why it is highlighted separately. By executing an A/A test, in which participants from the same control or treatment group are split and compared as if they were from different groups, the influence of non-human visitors can be detected.

Validation of the experiments executed as part of this thesis is done through a set of offline analyses carried out after the experiments’ execution. Further checking of the validity of the experiments is done implicitly through their design: By implementing two control groups an implicit “live A/A test” is part of the results gathered. At the same time, the results of the two control groups allow several other conceptual assumptions to be confirmed.

3. Experimental Study

The problem faced by the MastersPortal.eu website is that it does not succeed in convincing enough visitors referred by Google to stay on its website. A potential solution to this problem is the implementation of a recommender system for Master's programmes. Based on the introduction and previous background discussion an experimental study is proposed.
This study aims to provide insights into the effects of implementing a recommender system for Master's programmes on the MastersPortal.eu website. The study further aims to provide directions for the future use of recommender technologies within MastersPortal. This chapter defines the goals and overall design considerations of this experimental study.

3.1 Goal

Based on the problem description and the proposed solution discussed in *Chapter 1.3* the goal of the experimental study executed during this thesis is defined as follows:

> Determine *if* and *how* a recommender system can be implemented on the MastersPortal.eu website to reduce the bounce-rate of visitors referred by Google.

During the experimental study we aim to determine if implementing a recommender system has a positive, reducing, effect on the bounce-rate of visitors referred by Google. Apart from determining the overall effects of implementing a recommender system, we also hope to shed light on which approach to implement such a system is best suited for MastersPortal.
The experimental study is made up of two separate experiments serving as the data gathering steps, combined with multiple analyses on the data gathered.

The first experiment aims to test the effectiveness of the two main recommender approaches found in literature. These two approaches are compared with a baseline recommendation and two control groups to reliably determine their effects on the bounce-rate.

First Experiment: Compare the effects of implementing a content-based and collaborative recommender with a recommendation baseline and two control groups.

Once the effects on the bounce-rate of both the content-based and collaborative recommender are determined, a further analysis is executed to gather a better understanding of the factors that influence this performance. This further investigation is based upon a detailed analysis of contextual factors relevant to visitors of the MastersPortal.eu website.
Each contextual factor can be viewed as an additional dimension in the information need of a visitor. An understanding of these contextual factors is thus of importance to further optimising recommender performance. By classifying visitors based upon their contextual factors we hope to find additional indicators of their preference towards one of the recommender approaches evaluated during the first experiment.

To verify both the effects of implementing a recommender system and the contextual influences identified by looking at the different contextual visitor categories, a second experiment is executed.

Second Experiment: Confirm the influence of contextual factors on the performance of both the content-based and collaborative recommender approach, paying special attention to the geographic origin of each visitor.

Apart from confirming the effect each contextual factor has on the performance of the different recommender approaches, the second experiment also serves as verification towards the results of the first experiment.
Based on the results of the both experiments conclusions are drawn on the effects of implementing a recommender system on the MastersPortal.eu website. These conclusions are made more specific by taking into account the contextual factors identified in both experiments.
Based upon these conclusions an advice is provided to MastersPortal on how to be best implement a recommender system and how to improve this system in the future. As a final part of this thesis future direction for the MastersPortal.eu website concerning both recommender systems and more general topics are discussed.

In light of the results gathered from the experiments, one of the future directions identified in literature for MastersPortal's recommender system appears to be particularly relevant. In order to provide a more tangible future perspective for MastersPortal an extra analysis was added to this thesis.
The aim of this analysis is to determine if adaptive features can further increase the performance of the recommender system on the MastersPortal.eu website. This adaptive recommender system would utilise the contextual factors identified in this thesis to create visitor categories. Based on these categorisations an adaptive system can, without human intervention, come to a more precise understanding of each visitor's information need. An adaptive system can thus be used to automatically create a better match between the recommendations provided by the different recommender approaches and the interests of each visitor.
The final analysis provides a future perspective for MastersPortal grounded in the actual data gathered during the experiment. As such it offers a concrete direction for further developments within MastersPortal.

3.1.1 Evaluation Criterion

To judge the effects of the recommendation approaches implemented, an evaluation criterion is required. This criterion is the basis for all conclusions drawn with regard to the performance of the recommender system. By defining a single criterion at the start of this experiment we prevent ambiguity while analysing the results and drawing our conclusions.
The bounce-rate, as defined in *Chapter 2.4.2*, is used as the evaluation criterion for this experimental study. It has a clear effect on MastersPortal's bottom-line and can be easily measured within the confines of the experimental study. We define the bounce-rate ($BR_{Google\ ,R}$) as the fraction of visits referred by Google, exposed to a certain recommender approach R, that bounce.

$$BR_{Google\ ,R} = \frac{bounced\ visits_{Google\ ,R}}{total\ visits_{Google\ ,R}}$$

During this experimental study, bounced visits are those visits during which only a single page is viewed. These visits are considered not desirable with respect to MastersPortal's goals; in order to increase its revenue MastersPortal's visitors need to view more than a single page.

3.2 Assumptions

Based on the background discussion in the previous chapter, the experiment's goals and its evaluation criterion, two important assumptions emerge. This section discusses these two assumptions and their effect on the results of this thesis.

3.2.1 Bounce-Rate versus "Precision at k"

An important basis for our choice of the bounce-rate as the evaluation criterion for our experimental study is the assumption that a more relevant recommendation leads to a lower bounce-rate. A lower bounce-rate is thus assumed to correspond to a higher precision at k.
This assumption has no effect on the conclusions of the experiment. Within this experimental study we focus solely on the MastersPortal.eu website. MastersPortal hopes to increase its revenues by reducing the bounce-rate of its visitors. The goal of the recommender system on the MastersPortal.eu website is thus to reduce the bounce-rate, not to increase the relevance of the recommendations provided. If lowering the bounce-rate is achieved by providing less relevant recommendations that is fine as far as the current experiment's goals are concerned. This assumption is thus not further evaluated during this thesis.

In the remainder of this thesis, whenever the performance of a recommender system is discussed it can be assumed the "relevance towards the goals of MastersPortal" is meant; we do not refer to the formal definition of performance from an information retrieval perspective as presented at the start of *Chapter 2.4*.
This assumption is something which can be re-evaluated in a future study. If the assumption is confirmed, it will allow the results of this thesis to be further generalised. A strong correlation between the bounce-rate and the recall and precision of a recommender system entails the bounce-rate can be used to estimate, or evaluate, the performance proper of a recommender system.

3.2.2 Implicit Information Need

The main goal of the recommender system on the MastersPortal.eu website is to retain visitors referred by Google. The recommender system can thus not base its recommendations upon previously gathered information about a visitor; most of the visitors visit for the first time. The sole input of the recommender system is thus the programme currently viewed by the visitor. In our work we thus make an assumption that the Master's programme initially viewed by a visitor is a good indication of this visitor's information need.
By basing our recommendation decisions on the programme a visitor is referred to by Google, we essentially create a simple *pseudo-relevance feedback* system. Within pseudo-relevance feedback the top *k* documents retrieved by a user query are considered relevant to the user. In case of MastersPortal, the visitor "retrieves" a single document through a query on Google. Pseudo-relevance feedback in our case thus entails assuming the Master's programme retrieved through Google is relevant to the visitor.

In case of a collaborative recommender approach, violations of this assumption are not necessarily damaging to the relevance of the recommendations provided. Apart from positive relevance feedback, negative relevance feedback is also possible: If visitors with a certain information need are consistently referred to the same irrelevant programme, and some of these visitors manage to find a relevant programme through other means, a collaborative recommender approach can pick up on

this negative relation. Consequently, the collaborative approach will start offering the relevant programme as a recommendation to other visitors referred by Google to the irrelevant programme.

Considering the content-based recommender approach, violations to the assumption offer more of a problem. In the final chapters of this thesis we focus on the effects of a visitor's contextual factors. The conclusions of these chapters do indeed provide evidence of the fact that the above assumption can cause problems for the content-based recommender approach, which follows a *query-by-example* retrieval paradigm. Although the assumption is required to execute the experiment, it likely has an impact on its results.
Although the conclusions question the assumption, at the same time they provide means to overcome it. Future enhancements to MastersPortal's recommender system taking into account these conclusions are thus implicitly be able to better cope with the assumption. More specifically, by gathering information on a visitor's contextual categories, we can come to a better understanding of this visitor's needs.

3.3 Control and Treatment Groups

To evaluate the performance of the recommender system on the MastersPortal.eu website the experimental study consists of five participant groups. Based on the background discussion in *Chapter 2.2*, three treatment groups are defined. Furthermore, based on the discussion on experiment validation in *Chapter 2.5*, two control groups are defined.

Each treatment group consists of a technologically different recommender approach. The two control groups provide an overall baseline which is used to both verify the results of the experiments and to validate the assumptions concerning its controlled nature.
The first treatment group is considered to be the recommendation baseline. It provides an indication as to how the quality of the recommendations provided by the other two treatment groups influences performance.

1. Control "None"
   The first control group is the situation in which no recommendation is present. This is similar to the previous situation on the MastersPortal.eu website.
2. Control "Random"
   The second control group consists of a fully random set of programmes. This recommendation represents the situation in which an irrelevant recommendation is provided. Its goal is to measure the effect accomplished by the visual change to the website, without the actual added value of a relevant recommendation.
3. Treatment "Baseline"
   The first treatment is a recommendation based upon the hierarchy of academic disciplines present in the MastersPortal database. The academic disciplines create a hierarchy of Master's programme based on their academic content. This treatment is considered to be the baseline performance for any recommender system implemented on the MastersPortal.eu website.
4. Treatment "Content"
   The second treatment ranks programmes using a content-based recommendation approach and a tf-idf vector space comparison procedure. The unstructured programmes' descriptions, relevant to the academic contents of each programme, are used as basis for this recommendation.
5. Treatment "Collaborative"
   The third treatment is a collaborative approach that ranks programmes based upon the historical preferences of visitors to the MastersPortal.eu website.

An expectation implicit to the above definition of treatment groups is that the recommendation provided by *Treatment "Baseline"* performs significantly worse than those provided by either *Treatment "Content"* or *Treatment "Collaborative"*.
The recommendations provided by the baseline recommender do not utilise the unstructured information on the academic contents of the Master's programmes. As this is what is of real interest to visitors of the MastersPortal.eu website, my expectation is that the baseline recommendation will, on average, not be able to provide a recommendation as relevant as those provided by either the content-based or collaborative recommender approaches.

3.4 Roadmap

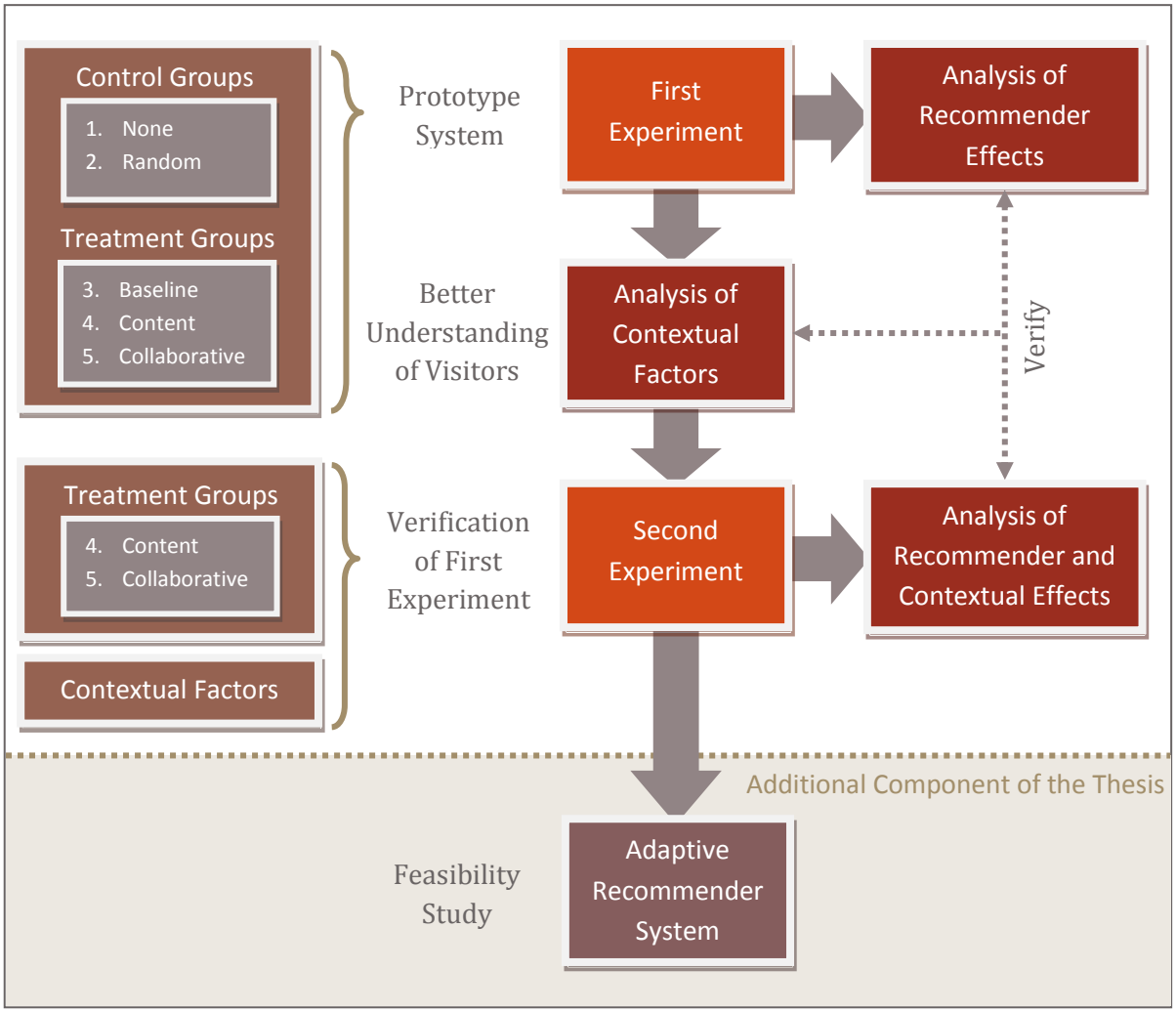


Figure 8: Roadmap of the Experimental Study

4. Recommender Experiment

Many visitors to the MastersPortal.eu website start their visit directly at a "programme details page". These pages, containing detailed information on a single Master's programme, are referenced from external search-engines such as Google.
Analysis of visitor behaviour shows that 84% of visitors arriving to such a programme details page from Google bounce back immediately after viewing this page. As a result, a lot of potential visitors are lost. In an attempt to decrease the *bounce-rate* the implementation of a recommender system is proposed.

This chapter describes the experiment executed to test the effects of implementing a recommender system on the MastersPortal.eu website. This is the first experiment that was executed as part of this thesis.

4.1 Experiment Design

The goal of the experiment is to determine the effect of several recommender systems which could potentially be implemented on the MastersPortal.eu website. The effect of each recommender system is judged by looking at the overall bounce-rate of visitors who, during their visit, are presented with this respective system.
The bounce-rate of a group of visitors is the fraction of their *visits* that consists of only viewing a single page. Note that each visitor, over a certain period of time, can have multiple *visits* to the MastersPortal.eu website.

The experiment takes the form of a controlled experiment in which as many factors as possible, apart from the recommender approach, are kept under control. Although the MastersPortal.eu website changes constantly, its state is kept as stable as possible during the execution of the experiment. No modifications to its lay-out are made and no new functionality is deployed over the course of the experiment. As such, the effects of factors other than the recommender approach on the results of the experiment are reduced as much as possible.
To compare the multiple alternatives, the experiment consists of series of hypothesis tests. Participants are equally divided amongst the potential recommender systems. Each participant is always presented with the same recommender approach. The visual aspects of each recommender approach are similar. Apart from differences in the technical approach of the recommender, and as such the set of programmes recommended, each participant sees exactly the same.
By varying this single factor we hope to determine the effect each recommender system has on the bounce-rate. Through this measure it is possible to determine the performance and relative effectiveness of each of the recommender system as implemented on MastersPortal.eu.

All five treatment and controls groups as discussed in *Chapter 3.3* are taken into account during the first experiment. Additional operational details on each of their implementations are provided in *Chapter 4.3* and in *Appendix D*.

4.1.1 Conceptual Model

The overall design of the experiment is transferred into a conceptual model, which is provided below in *Figure 9*. The diagram in this figure provides a high level overview of the flow and results of the experiment.

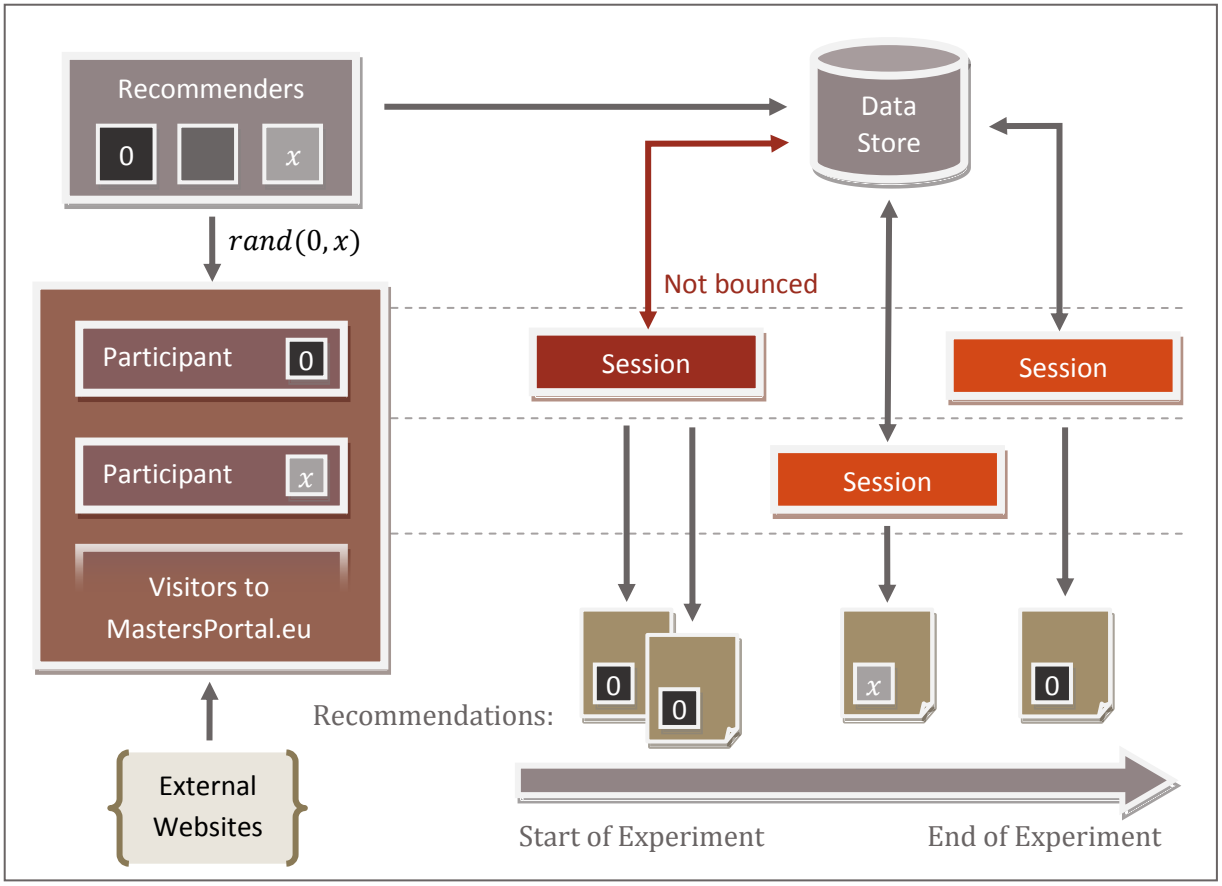


Figure 9: Conceptual model of the first experiment.

Only participants referred to MastersPortal.eu through external websites are relevant for the experiment. The conceptual model in *Figure 9* highlights two hypothetical participants. For these two participants the bounce-rate equals $\frac{2}{3}$, as only one out of the three sessions, or visits, generated by the participants consists of more than a single page view.

A session as presented in the diagram above is defined as a set of requests by a visitor with at most sixty minutes in between them. After this interval expires, any subsequent requests will be considered a new session.
This definition is very similar to the "visit" definition used in the introduction to this thesis. For the remainder of the thesis, the terms "session" and "visit" might thus be used interchangeably.

4.2 Hypotheses

Based on the experiment design discussed in the previous chapter, three hypotheses are formulated for the first experiment.

Hypothesis 1: Providing an irrelevant recommendation has no effect on the bounce-rate.

Hypothesis 2: Providing a relevant recommendation decreases the bounce-rate.

Hypothesis 3: Providing a more relevant recommendation than the baseline recommendation further decreases the bounce-rate.

Details on the exact statistical procedures utilised are provided in *Chapter 2.5.1*.

4.2.1 Hypothesis 1

For the first hypothesis the current situation on the MastersPortal.eu website, without a recommendation present, is compared against the situation where a fully random recommendation is presented. This leads to the following hypothesis:

Hypothesis 1: Providing an irrelevant recommendation has no effect on the bounce-rate.

H0: The bounce-rate for *Control "None"* and *Control "Random"* are equal;
H1: The bounce-rate for *Control "None"* and *Control "Random"* are not equal;

The underlying assumption for this hypothesis is that a fully random recommendation will be perceived by visitors as being completely irrelevant. The goal is to determine whether the mere presence of an element labelled "recommendation" has an influence on the visitors.
My expectation is that the mere presence a "recommendation" will not convince a significant amount of visitors to stay on the MastersPortal.eu website instead of returning to their original referrer.
Apart from validating the above assumption, the initial hypothesis also functions as an A/A testing procedure, as discussed in *Chapter 2.5.2*. We expect both control groups to perform similarly. If it turns out this is not the case there could be issues with the implementation and execution of the experiment itself.

4.2.2 Hypothesis 2

The second hypothesis compares the current situation on MastersPortal.eu, no recommendation, against three possible recommender systems. All three recommender systems contain a technologically different implementation. To properly test this hypothesis, it is split up into three sub-hypotheses.

Hypothesis 2.1: Providing a recommendation based upon a hierarchical programme categorisation decreases the bounce-rate.

H0: The bounce-rate for *Control "None"* and *Treatment "Baseline"* are equal;
H1: The bounce-rate for *Treatment "Baseline"* is lower;

Hypothesis 2.2: Providing a content-based recommendation decreases the bounce-rate.

H0: The bounce-rate for *Control "None"* and *Treatment "Content"* are equal;
H1: The bounce-rate for *Treatment "Content"* is lower;

Hypothesis 2.3: Providing a collaborative recommendation decreases the bounce-rate.

H0: The bounce-rate for *Control "None"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Collaborative"* is lower;

The second hypothesis aims to determine whether including *any* relevant recommendation will lower the bounce-rate when compared to the situation where no recommendation is present. It serves to provide a baseline for the third hypothesis.

4.2.3 Hypothesis 3

The third hypothesis compares the performance of the baseline recommender based on a hierarchical programme categorisation, *Treatment "Baseline"*, against the two automated recommenders: *Treatment "Content"* and *Treatment "Collaborative"*. In order to properly test this hypothesis, it is again split up into three sub-hypothesis.

Hypothesis 3.1: The content-based recommendation decreases the bounce-rate further than the baseline recommendation.

H0: Bounce-rate for *Treatment "Baseline"* and *Treatment "Content"* are equal;
H1: Bounce-rate for *Treatment "Content"* is lower;

Hypothesis 3.2: The collaborative recommendation decreases the bounce-rate further than the baseline recommendation.

H0: Bounce-rate for *Treatment "Baseline"* and *Treatment "Collaborative"* are equal;
H1: Bounce-rate for *Treatment "Collaborative"* is lower;

Hypothesis 3.3: The content-based recommendation decreases the bounce-rate further than the collaborative recommendation.

H0: Bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: Bounce-rate for *Treatment "Content"* is lower;

The third hypothesis aims to determine how the technology used to generate the recommendations affects the bounce-rate. Specifically, do the automated and technologically more advanced recommenders lead to a lower bounce-rate? The main expectation for this hypothesis is that the two automated recommenders, *Treatment "Content"* and *Treatment "Collaborative"*, provide a superior recommendation compared to the baseline situation and as such lead to significantly lower bounce-rates.
This third sub-hypothesis of the final hypothesis aims to determine whether there is a difference between the two automated recommendations. The order of the hypothesis indicates my personal expectation: *Treatment "Content"* outperforms *Treatment "Collaborative"*.

4.3 Experiment Setup

This chapter focuses on the practical implementation of the experiment, details its execution and highlights all further assumptions and limitations introduced to the experiment based upon this implementation.

4.3.1 Recommender Approaches

On the detailed information pages of the MastersPortal.eu website, a "Related Programmes" box is presented to the visitor. This box contains eight programmes related to the Master's programme currently viewed. An example of the visual presentation of the "Related Programmes" box is provided in *Appendix C2*.

The number of programmes shown in the box is chosen such that for most visitors the entire "Related Programmes" box is visible upon loading the page, without the need for scrolling the window. The selection of this number is a judgment call, based upon a general experience with the distribution of screen resolutions used by MastersPortal.eu's visitors.

Results from the second experiment indicate that around 70% of visitors do indeed have a screen resolution large enough to see the entire recommender box at once. The impact of a visitor's screen resolution can also be seen in *Appendix C2*. Around 30% of visitors see less than the "Medium" screen resolution overlaid on the screenshot. These visitors were thus not able to evaluate the entire "Related Programmes" box without scrolling their browser window.

For the experiment, five groups of visitors are defined. Two control groups and three treatment groups. Below follows a short discussion of each approach.

The visual presentation of the "Related Programmes" box does not differ between the treatment groups and the *Control "Random"* group. The *Control "None"* group will of course not show anything to the visitor. The recommendations for all programmes are pre-computed before the start of the experiment. The speed of at which a recommendation is computed is thus equal for all recommendations throughout the experiment, irrespective of their approach.

Random "Recommender"

The random "recommender" randomly selects a set of Master's programmes from the 14.000 programmes present in the MastersPortal database. These programmes are subsequently presented as if they are relevant programmes. Apart from coincidental relevance, the programmes provided by the random recommender can be considered completely irrelevant.

In order to prevent any undue influences, a random "recommendation" for each programme is pre-computed and stored before the experiment starts. So, tough the recommendation is randomly generated, it will remain the same throughout the course of the experiment.

Baseline Recommendation

Each programme in the MastersPortal database is categorised into one or more academic disciplines. These disciplines are assigned to the programmes by their administrators. The academic disciplines assigned to each programme provide a human judgement on the optimal academic hierarchy of all Master's programme. An overview of the disciplines is provided in *Appendix K*. Each programme has from one and five disciplines assigned to it.

The recommender looks at the disciplines assigned to the reference programme and bases its recommendation on discipline overlap with other programmes in the database. Programmes that share the largest amount of disciplines are considered to be most related. Due to the limited number

of disciplines per programme, the distinctive power of this recommendation is lower than that of the two automated recommenders discussed next.

Content-Based Recommender

Each Master's programme in the MastersPortal database contains a detailed description outlining the programmes goals, main subjects and requirements. The content-based recommender looks at the detailed description of the reference programme and bases its recommendation of related programmes on textual similarities. Textual similarities are computed using a *tf-idf vector space comparison* approach. In *Appendix D1* a further overview of the implementation of this recommender is provided.

Collaborative Recommender

From the historical browsing habits of previous visitors to the MastersPortal.eu website, groups of related programmes can be harvested. When a visitor, during a single visit, views a set of Master's programmes, these programmes are assumed to be related and filling the visitor's information need. The collaborative recommender bases its recommendation on these groupings, recommending programmes which occur frequently in the same group as the reference programme. In *Appendix D2* an overview of the harvesting and filtering procedure of historical data is provided.

4.3.2 Execution of the Experiment

All visitors to the English-language version of the MastersPortal.eu website participate in the experiment. The recommendation itself is presented on all pages which provide detailed information on a single Master's programme. All *visits* referred by Google to one of these pages are taken into account while compiling the results of the experiment. All other visits are ignored.

Running the experiment on the live MastersPortal.eu website adds some additional complexities, but there is no feasible alternative that allows the experiment to be executed in a more controlled environment. The experiment focuses on converting visitors referred by Google and as such can only be executed there where these visitors actually arrive. Automatically transferring a part of these visitors to an experimentation website with reduced functionality (c.q. a more controlled environment) would degrade their experience too much.
The experiment runs for four weeks. The visitor statistics trend for the months leading up the experiment suggests that this interval is more than sufficient to provide high statistical power to the results of the experiment.

Each new participant is randomly assigned into one of the five experimentation groups**.** Once assigned, the participant is kept in this group for the duration of the experiment. Each recommendation issued for the participant will thus be based on the same approach. This is an important factor to ensure the controlled nature of the experiment.
Cookies are used to enforce the group assignment of each participant. As not all participants accept cookies, logging code is put in place to ensure non-acceptance of cookies is detectable and can be acted upon.
From the results presented in *Appendix G1* we conclude that around 6% of the participants rejected the experiment cookie and were excluded from the experiment based on this determination. This number is sufficiently low to regard the approach of using cookies to enforce the controlled nature of the experiment as valid.

In *Appendix B* an overview of the data collection procedures employed during the experiment is provided.

Pre-filtering is applied to all participants and aims exclude the most common non-human "users" of the Internet. In case of the MastersPortal.eu website, *Googlebot*, *Yahoo Slurp* and *msnbot* together account for nearly 25% of all page views. Pre-filtered participants are not presented with a recommendation and are excluded from all further analyses to prevent any polluting influence they might have.
Before the results of the experiment are analysed, a second post-filtering step is applied to further prevent interference from invalid participants. Attempts are made to exclude all non-human visitors, all participants who did not accept the experimentation cookie and all participants who, at some point, viewed an incomplete recommendation.
The experimentation cookie is used to ensure each participant always views the same type of recommendation; participants who do not accept this cookie violate the constraints of the controlled experiment. The same goes for participants who, at some point during the experiment, were presented with an incomplete recommendation. *Figure 10* below provides a schematic overview of the entire filtering process.

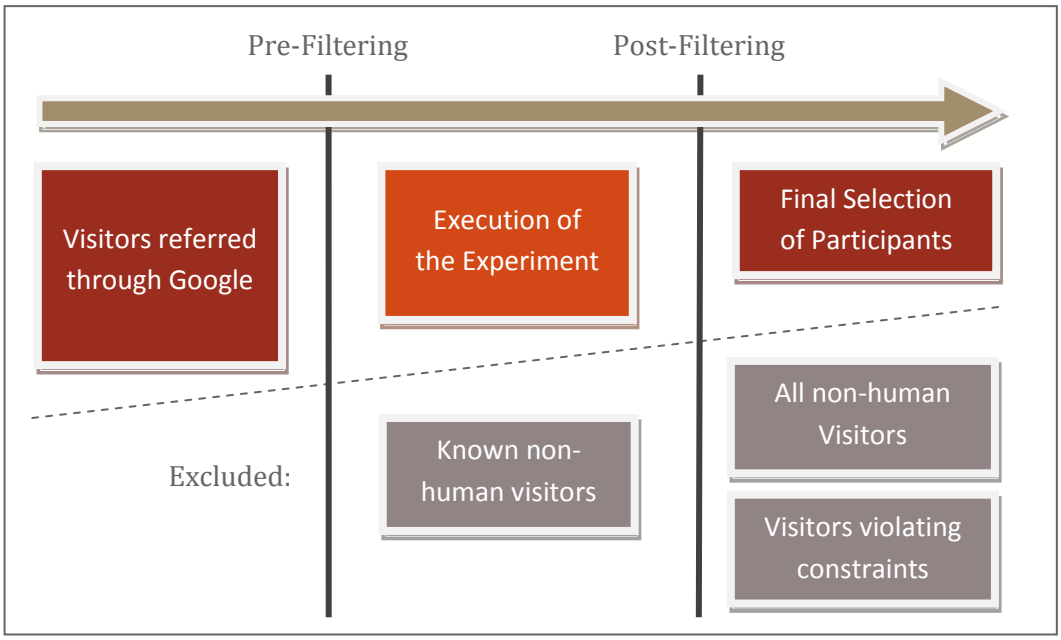


Figure 10: Schematic overview of the participant filtering procedure

Finally, the bounce-rate metric is setup as follows: The fraction of visitors in each experiment group who receive more than a single recommendation during their visit *or* click on an item in "Related Programmes" box. This way, the indirect effect of the recommender system is also measured, as discussed in *Chapter 2.4.2*.

4.4 Results

The recommender experiment was executed from January 14th of 2010 up to and including February 11th of 2010 on www.mastersportal.eu. Over the course of the experiment, its progress was monitored through a simple live environment, designed to catch problems with the experiment and its data collection early. The final results, as presented below, were generated through an offline analysis executed after data collection had completed.

4.4.1 Results Overview

The daily number of participants created over the course of the experiment is displayed in *Figure 11* below. The first four and last four days of the experiment are considered as warm-up and cool-down periods, leaving three full weeks, ranging from Monday to Sunday, for the actual experiment. The experimentation interval thus runs from January 18th of 2010 up to and including February 7th of 2010.

During the first few days of the experiment some minor tuning was done to optimise performance of the recommenders and to correct several minor issues with logging of the results. It was decided to drop the first few days of data from the experiment in order to prevent these minor issues from influencing the results of the experiment.
The reason for leaving out several days at the end lies in the fact that it allows for participants with multiple sessions to also be counted as such properly near the end of the experiment. The effects of this decision are discussed in more detail with regard to *Figure 13*, further down.
The experimentation interval was constructed by fully excluding all participants that were seen before the 18th. Thus, the return visits of a participant that was first seen before the 18th, but occurred after the 18th, were also excluded.

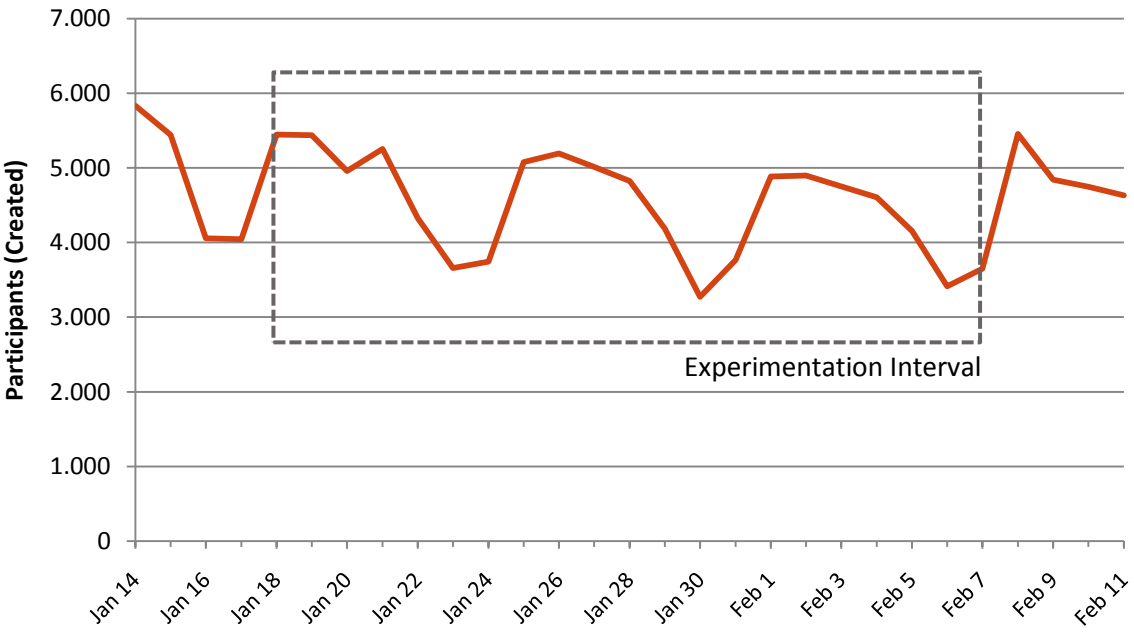


Figure 11: Daily number of participants created during the experiment, experimentation interval outlined

For the complete duration of the experiment, 133.558 participants were recorded. Taking into account the three week experimentation interval 94.513 participants remained. In *Figure 11* the experimentation interval is outlined by the grey dotted box.

After selection of the experimentation interval, rigorous filtering was applied to the participants inside the interval. The graph in *Figure 12* provides an overview of the number of participants before and after filtering. A detailed description of the filtering executed is presented in *Appendix G1*.

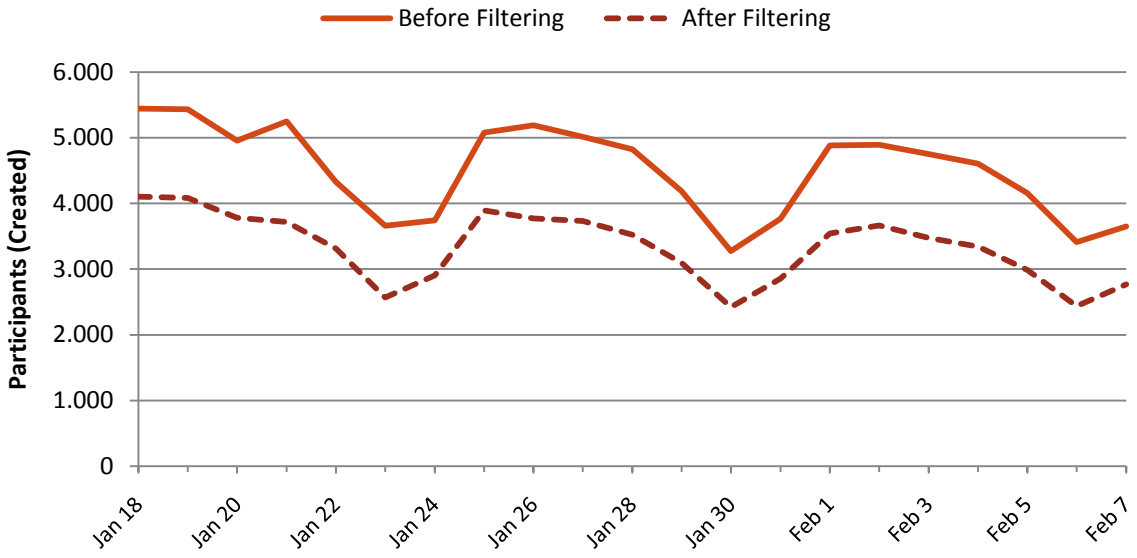


Figure 12: Daily number of participants created during the experimentation interval, before and after filtering.

After the filtering step the final selection of participants was completed. The next step was to apply a selection procedure to the sessions generated by the participants.

As noted in *Chapter 4.3*, only sessions which started with a referral from Google are considered relevant for the experiment. The graph in *Figure 13* provides an overview of the total number of sessions generated by our filtered participants and subsequently the number of sessions after those not starting from Google were excluded.

Note that a single participant can have multiple sessions, of which not all necessarily stem from Google. In these cases, the participant *is* considered valid, but only its sessions that started by a referral through Google are included.

This final selection step leaves us with 58.853 sessions, representing 55.230 participants, of which some might thus not have had all their sessions included.

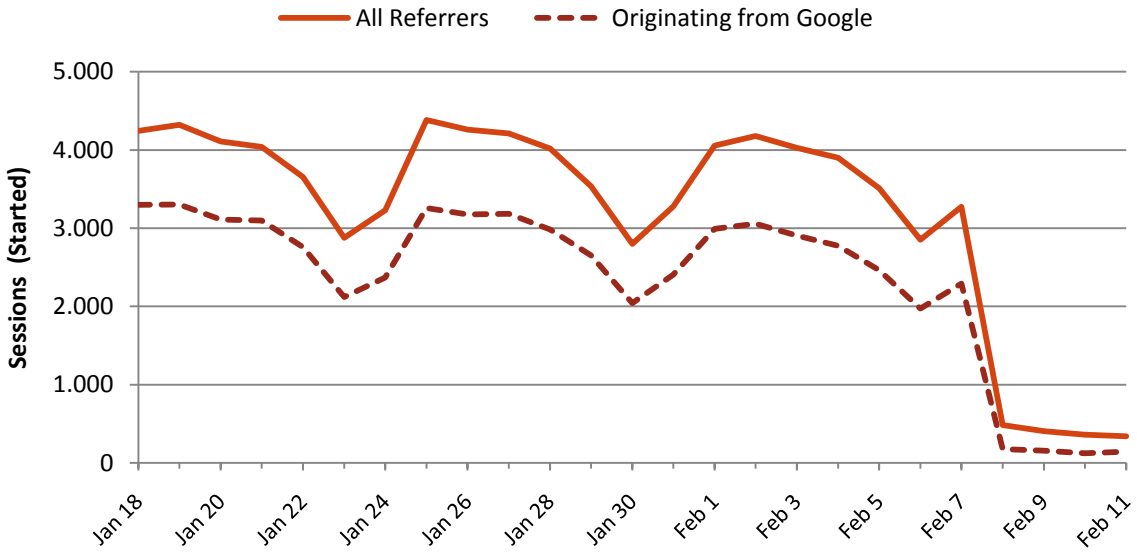


Figure 13: Daily number of sessions started, overall and originating from Google, during the experimentation interval

The dip at the end of the graph in *Figure 13* is caused by limiting the experimentation interval as explained earlier. On the 7th of February, the addition of new participants to the offline analysis ceased. Sessions of existing participants were still counted and thus some additional sessions were recorded.
The number of sessions added this way is rather low, in total around 1.000. Initially it was not entirely clear how large the effects of recurring visits would be. In hindsight, making this further limitation was not absolutely necessary. In light of the fact that more than enough data is gathered during the experiment, no attempts were made to re-evaluate this decision.

A final step in the analyzing the results of the experiments is looking at the distribution of participants amongst the five experimentation groups. Participants should be distributed equally amongst the five groups. An overview of the grouping is presented in *Figure 14*.

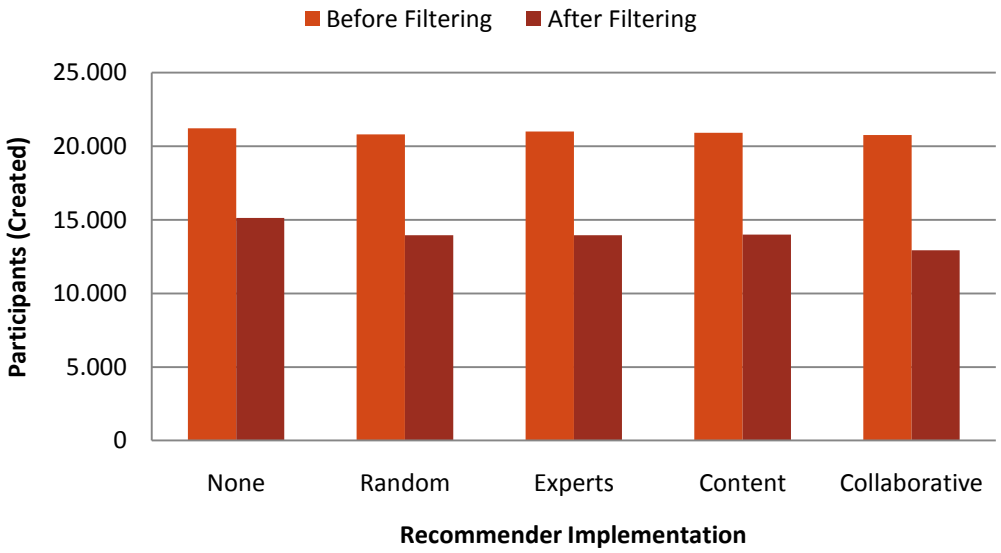


Figure 14: Total number of participants for each of the experiment groups.

The initial distribution of participants amongst the five groups is completely fair; both illustrated by the diagram in *Figure 14* above and the data in *Table 1* below. We do see that after applying the participant filter, the distribution becomes less balanced.

Participants	[Before Filtering]		[After Filtering]	
None	21.230	20,3%	15.128	21,6%
Random	20.797	19,9%	13.953	19,9%
Baseline	20.996	20,1%	13.961	19,9%
Content	20.923	20,0%	14.007	20,0%
Collaborative	20.771	19,8%	12.933	18,5%

Table 1: Distribution of participants amongst the five experiment groups

The collaborative recommendation appears to be more prone to providing incomplete recommendations. As such, it has substantially more participants excluded through filtering. Although at first sight this seems a violation of the experiment assumptions, it is in effect only a complicating factor in the analysis of the results.
The initial distribution of participants amongst the groups is fair. Only after applying rigorous filtering an imbalance is introduced. If we assume the filtering within the groups is completely random, this imbalance should not influence the overall validity of the results. Evidence of this assumption is provided by *Figure 49* in *Appendix G1*.

4.4.2 Multiple Comparison of Means

After the final selection of sessions was made, the bounce-rate for each of the five groups within the experiment could be computed. In this chapter, the bounce-rates are interpreted using a multiple comparison procedure. The goal of this analysis is to provide a ranking amongst the five groups within the experiment.

As the bounce-rate is a binomial variable in this experiment, it is approximated normally for the purpose of applying the multiple comparison procedure. In applying the multiple comparisons, Bonferroni's method is used. This method makes neither an assumption of equal sample size nor of equal variances between the groups. Statistical significance is set to 99% for all analyses in this chapter. The details of why this multiple comparison procedure was chosen and the exact testing procedure are outlined in *Chapter 2.5*.

Due to the rigorous participant selection, the number of participants between the five groups is not equal. As discussed earlier, this does not invalidate the results, but it does complicate the interpretation of the results of the multiple comparison procedure somewhat. As such it was decided to assume an equal number of participants for each group.
The group sizes are assumed to be 10.098 for each group**,** equal to the lowest overall number of participants in a group. Note that this change has no effect on the actual bounce-rates, merely on confidence intervals generated by the multiple comparison procedure.
In *Appendix E* an overview of the same multiple comparison procedure is provided without the above assumption. The results in this appendix indicate the assumption made does not cause any problems with the interpretation and validity of the results.
As a result of this secondary analysis, a somewhat ambiguous situation arises around the difference in performance between the content-based recommender and the collaborative recommender. This situation will be discussed in more detail further down.

In *Table 2* an overview of the results of multiple comparison procedure is presented. It shows the groupings inferred from the tests' results. As can be seen in the table, three distinct groups are found. Participant from the *Control "None"* and *Control "Random"* are considered to behave homogenously, participants from the groups *Treatment "Content"* and *Treatment "Collaborative"* behave homogeneously. Participants from *Treatment "Baseline"* form a group by themselves.

	Homogeneous Groups		
None	•		
Random	•		
Baseline		•	
Content			•
Collaborative			•

Table 2: Homogeneous groups within the first experiment's results

By combining the results from *Table 2* and *Table 3* we can come to a performance ranking of the five experiment groups, or more precisely, the three homogeneous groups identified.
The group consisting of the "none" and "random" recommender perform worst, with a bounce-rate of around 90%. The "experts" recommender has a bounce-rate of around 85,5%. The group consisting of the "content" and "collaborative" recommenders performs best and has a bounce-rate of around 82%.

	Recommendations	Bounces	Bounce-rate
None	12.540	11.322	90,29%
Random	11.789	10.620	90,08%
Baseline	11.787	10.070	85,43%
Content	11.829	9.639	81,49%
Collaborative	10.908	8.948	82,03%

Table 3: Group bounce-rates for the first experiment

Finally, in *Table 4* the contrasts between all experiment groups as computed through the multiple comparison procedure are provided. These contrasts provide the absolute differences, in number of participants bounced, between the groups and the related confidence intervals.
Significant differences are marked with an asterisk. As the number of participants in each group is assumed to be equal, the confidence intervals are equal for all the group contrasts too.

Contrast	Difference		Confidence
Baseline – None	-530,0	*	59,78
Baseline – Random	-507,0	*	59,78
Baseline – Content	430,0	*	59,78
Baseline – Collaborative	371,0	*	59,78
None – Random	23,0		59,78
None – Content	960,0	*	59,78
None – Collaborative	901,0	*	59,78
Random – Content	937,0	*	59,78
Random – Collaborative	878,0	*	59,78
Content – Collaborative	-59,0		59,78

Table 4: Group contrasts for the first experiment (* indicates statistically significant difference)

The results in *Table 4* show that the difference between the "none" and "random" control groups is indeed very small; well within the 99% confidence interval for the test.
The results for the content-based and collaborative recommenders indicate the lack of significance

here is a border case. The absolute difference is only $0{,}78$ participant away from the 99% confidence interval. Reducing the confidence interval to 98% provides a statistically significant difference between both approaches. This situation will be looked at in more detail in the subsequent chapter on *Validation*, before a final conclusion can be drawn regarding the hypotheses.

Based on the experiment results presented above, conclusions concerning the hypotheses of the experiment are discussed in *Chapter 4.6*. Before these final conclusions are drawn, an extensive validation of the results is performed in *Chapter 4.5*.
Firstly we will quickly revisit the visitor statistics presented in *Chapter 1.2*. Some differences between the visitor statistics presented in the introduction to this thesis and the results of the first experiment are observed. The cause of these differences is discussed in the next section.

4.4.3 Comparison with Initial Visitor Statistics

Comparing the overall results of the first experiment with the visitor statistics as presented in *Chapter 1.2* we notice a substantial discrepancy. The results as presented in the introduction to the thesis list an average bounce-rate of 84%. The results from the experiment discussed in this chapter indicate a bounce-rate of 90% under similar conditions.

A short investigation reveals that the discrepancy is for a large part caused by the less extensive filtering applied to visitor statistics presented in the introduction. Especially the behavioural filtering component discussed in *Appendix F* is a major influence in this respect.
A further factor is the stateless nature of the HTTP protocol. This makes it difficult to gather all information required from the web server log-files. Some information has to be inferred and as such, the harvested results are only indicative of actual performance. They will always deviate somewhat from performance as measured by the experiment.

Apart from these two points, the MastersPortal.eu website received a complete redesign in between the analysis executed as part of the introduction to this thesis and the execution of the first experiment. It is possible that this redesign had an unwanted, negative, effect on the bounce-rate.

Should future investigations indicate the above explanation is insufficient the bounce-rate has only risen since the motivation for this thesis was provided. Therefore, the validity of the conclusions drawn during the introduction to this thesis is not in doubt.

4.5 Validation

This chapter provides an overview of the validation performed on the results of the first experiment. The full results of the validation are available in *Appendix G1*.

For the first experiment, two validation steps were performed. Firstly, the results of the experiment were compared against results generated by the third-part statistics package Webalizer. Secondly, a purpose-build web server log-file harvester is used to validate the results of the experiment.

From this first validation step no strange results are observed. Both the visitor statistics provided by the experiment and the visitor statistics provided by Webalizer correlate strongly. There are no strange spikes in visitor behaviour in either of the datasets.
Secondly, the results of the experiment were compared against statistics generated by the "MastersPortal harvester" discussed in *Appendix A*. This purpose-build harvester retrieves the number of sessions started and the number of recommendation clicks from the web server log-files. Although harvesting web server log-files only provides an approximation of the actual results, this approximation should be closely related to the results of the experiment. Looking at the results in *Figure 50* and *Figure 51* of *Appendix G1* we see the approximation closely matches the result of the experiment. No spikes or major deviations are present.

Concluding we can state that the results of the first experiment are confirmed by two independent analyses, constructed utilising a secondary data source. The validity of the experiment's results is thus confirmed.

The next section of this chapter briefly discusses the stability of the experiment's results over time. This is a validation issue which arose during the analyses of the results of the first experiment.

4.5.1 Stability of the Results

As a result of the discussion in *Chapter 4.4*, attention was drawn to the fact that the performance difference between the content-based recommender and the collaborative recommender is rather small.
The "Content" treatment has a $0{,}5\%$ lower bounce-rate; this difference is not statistically significant. It is somewhat of a border case though, as slightly relaxing the statistical parameters does lead to the conclusion that the "Content" treatment performs better: At the 99% confidence level both recommenders perform equal; at the 95% confidence level the "Content" treatment outperforms the "Collaborative" treatment.

In order to gain a better insight in the relation between the two recommendation approaches, they are graphed over time in *Figure 15*.

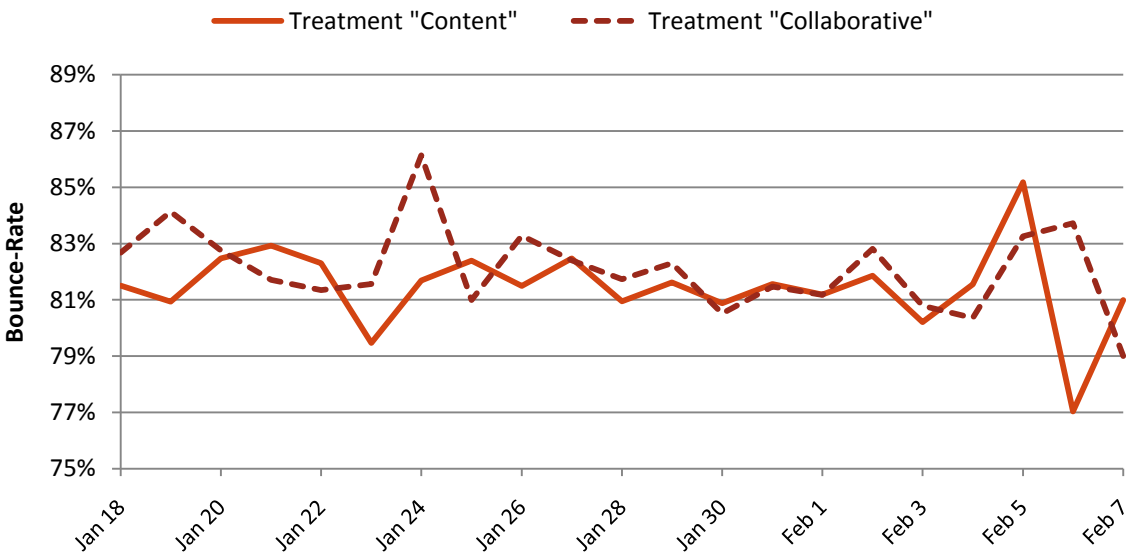


Figure 15: Daily performance of the "Content" and "Collaborative" treatments during the first experiment.

From the graph in *Figure 15* it is clear that the performance of the "Content" and "Collaborative" treatments is close. Strong conclusions on whether one treatment consistently outperforms the other are quite difficult when taking the above graph into account. The "Content" treatment does perform slightly better, but the biggest differences seem to be caused by a single upward spike for the "Collaborative" treatment and two downward spikes for "Content" treatment.

4.6 Conclusions

Based upon the results and validation of the experiment discussed in the previous sections, the following conclusions are drawn with relation to the hypotheses posed. An overview of the results of the hypotheses is presented in *Table 5* below.

	Result of Hypothesis Test	Expectation
Hypothesis 1	**H^0 Not Rejected**	**Do not Reject H^0**
Hypothesis 2	**-**	**-**
Hypothesis 2.1	H^0 Rejected	Reject H^0
Hypothesis 2.2	H^0 Rejected	Reject H^0
Hypothesis 2.3	H^0 Rejected	Reject H^0
Hypothesis 3	**-**	**-**
Hypothesis 3.1	H^0 Rejected	Reject H^0
Hypothesis 3.2	H^0 Rejected	Reject H^0
Hypothesis 3.3 *	H^0 Not Rejected	Reject H^0

Table 5: Overview of results for hypotheses 1, 2 and 3; * indicates result not matching expectation

Hypothesis 1: Providing an irrelevant recommendation has no effect on the bounce-rate.

For the first hypothesis, the null hypothesis cannot be rejected, indicating there is no statistically significant difference in performance between the situation with no recommendation and the situation with a random recommendation.
Looking at the results presented in *Table 3* and *Table 4* we see strong evidence towards this conclusion. The difference in bounce-rate between the two control groups is 0,21%. As a result we conclude that providing an irrelevant recommendation to a visitor of MastersPortal.eu who is referred by Google has no substantial effect on the bounce-rate of this visitor.

Hypothesis 2: Providing a relevant recommendation decreases the bounce-rate.

For the second hypothesis we reject the null hypothesis of all three sub-hypotheses. This provides statistically significant evidence that providing a relevant recommendation lowers the bounce-rate for visitors of the MastersPortal.eu website who are referred by Google.

Hypothesis 3.1 & 3.2: Providing a more relevant recommendation than the baseline recommendation further decreases the bounce-rate.

For the first two sub-hypotheses of the third hypothesis, the null hypothesis is rejected. Both the content-based and the collaborative approach outperform the baseline recommendation for visitors referred by Google.

Hypothesis 3.3: Providing a content-based recommendation decreases the bounce-rate further than providing a collaborative recommendation.

For the third sub-hypothesis of *Hypothesis 3*, the null hypothesis cannot be rejected. This leads to the conclusion that the content-based recommender approach and the collaborative recommender approach perform equally well.
This conclusion is though somewhat of a border case. If the confidence interval for the hypothesis test is lowered from 99% to 98%, the content-based recommender approach does gain a statistically significant performance lead.

In the validation discussion of *Chapter 4.5.1*, additional attention is paid to the close proximity of the performance of the content-based and collaborative recommender approaches. Considering *Figure 15* from this chapter it becomes clear that drawing a strong conclusion will be difficult.
Adding to the uncertainty is the additional analysis executed in *Appendix E*. This appendix clarifies an important assumption made while analysing the hypothesis tests. In doing so, it provides a further indication that the content-based recommender approach does perform significantly better than the collaborative recommender approach.

The next chapters of this thesis look further into the relative performance of both the content-based and collaborative recommender approaches. Through the analyses in these chapters we will attempt to come to a more definitive conclusion. For the time being though, the content-based and the collaborative recommender approaches are considered to perform equally well.

5. Contextual Factors

This chapter describes a post-analysis executed on the results of the first experiment. The goal of this analysis is to determine what the effects on recommender performance are if we classify visitors based contextual information.
The contextual effects provide additional insights regarding the performance difference between the content-based and collaborative recommenders. Furthermore, the analysis provides grounding for a further discussion on an adaptive recommender system which takes into account this contextual information.

5.1 Introduction

The experiment detailed in the previous chapter indicates that providing recommendations to visitors referred by Google significantly decreases their bounce-rate. Both the content-based recommendation as well as the collaborative recommendation yield good results. Although no hard statistical evidence could be provided, there is a clear indication that the content-based recommendation outperforms the collaborative recommendation.

In an attempt to shed more light on the relative performance of these two recommenders, a post-analysis is executed on the data gathered during the first experiment. Within this post-analysis, we again look at the performance of both recommenders, but this time we classify participants based upon several contextual factors.
The goal of this analysis is to see if the relatively close performance of both the content-based and collaborative recommenders is a property equal for each visitor, or if a difference can be observed when we classify visitors into different categories based upon contextual factors.

The analysis is executed by taking each of the 58.853 sessions as discussed in *Chapter 3* and classifying them along the lines of several contexts. Subsequently, for each classification within each context, bounce-rates are computed and a conclusion on statistical significance is drawn. During this analysis we are mostly interested in the difference between the content-based and collaborative recommenders. Whenever interesting observations on the other recommenders are available, they are noted, but they are not taken into account any further.

A difference between the experimentation groups is considered statistically significant if the bounce-rates computed using Bonferroni's method of multiple comparisons differ at the 99% confidence level. This is the same statistical procedure as used in *Chapter 3*.
All further filtering conditions as outlined in *Chapter 3* and *Appendix F1* were also applied prior to the subsequent analyses. This includes the strict non-human visitor filtering *and* only taking into account visitor sessions that originate from Google.

5.1.1 Simpson's Paradox

By splitting the results of the first experiment into several subsets, the possibility of an erroneous conclusion caused by failing to spot Simpson's Paradox need to be taken into account. The conditions required to encounter this paradox are though not present in the analyses executed in this chapter.

The subsets constructed for the upcoming analyses are taken from the overall dataset with as goal to determine the performance of each of these subgroups individually. The overall results of the experiment as a whole have already been determined in the previous chapter and are not calculated by summing up the relative performance of the each of the groups discussed in this chapter. As such the overall result cannot be influenced by Simpson's Paradox.
The performance of each subgroup is compared to the other subgroups stemming from the same context. I am specifically interested in determining differences in performance amongst the groups that are *not* apparent from the overall result. As such, the cause of Simpson's Paradox is a factor in the analyses, but due to the way its goals are defined, the paradox poses no threat to the validity of their interpretation.

In the next section an overview of the different contexts is provided. Subsequently the results of the post-analysis are provided and the most important observations are highlighted.

5.2 Contextual Factors

As part of the discussion in *Chapter 2.3*, many potentially interesting contexts have been identified. Including all these contexts in the subsequent analysis would only serve to overly complicate it. As such, only four contexts are taken into account; this chapter details these contexts.
The contexts are selected on the basis that they are both easy to measure and relevant to the MastersPortal.eu website. The contexts focus on easily identifiable properties indicating the "interests" of a visitor.

1. Geographical Origin
2. Google Query
3. Academic Discipline
4. Screen Resolution

The four contexts above can roughly be divided into two groups. The first three contexts aim to identify an intrinsic "interest" of the visitor. They provide information on "what" the visitor might be looking for on the MastersPortal.eu website.
The final context refers to a more technical property of the visitor's experience on the MastersPortal.eu website: The screen resolution can influence a visitor's behaviour as it has an important impact on where the visitor's attention is drawn to.

The following sections shortly introduce each of the four contexts in relation to the MastersPortal.eu website and provide a short overview of their practical implementation. In the next chapter, the bounce-rate effects of dividing visitors along these contexts are examined.

5.2.1 Geographical Origin

The first context is centred on the geographical origins of the visitors. It aims to detect geographical differences in visitor behaviour. People from different parts of the world, from different cultures specifically, have different preferences. As such they might also show differences in performance when it comes to recommender approaches.
The geographical origin is based upon the location from where a visitor's initial request originates. Due to the distributed nature of the Internet this might not always be where the visitor is actually from, but a very strong correlation is to be expected.

The geographical location of the visitor is determined by comparing their IP address against the GeoIP-Lite database (MaxMind, 2009). This is a freely available information source linking each IP address in the world to a country, with a reported accuracy of 99,5%.
The visitor's continent of origin is determined by matching the country as previously detected against the list of countries in each continent defined through the United Nations' geoscheme (United Nations Statistics Division, 2009).

5.2.2 Google Query

The "Google" query context aims to split visitors based upon the query they entered on Google before they were referred by Google to the MastersPortal.eu website. As all visitors taken into account during the first experiment were referred by Google, most of them did indeed provide a search-query on Google. These queries potentially provide a great deal of information on the specific interests of each visitor.

A separation is made between visitors who, through their query, indicate a general interest in a Master's degree and visitors who do not. Within the former category a second subdivision is made.

In order to determine the query-type, a simple textual analysis is executed on each of the Google queries stored during the first experiment. An overview of the classification procedure is provided below. The exact procedure used to classify each of the queries can be found in *Appendix 1*.

INTEREST IN MASTER'S DEGREE: Indicates a general interest in the information offered on the MastersPortal.eu website. This condition is triggered when terms like "master" or "degree" are present in the query.

- LOCATION INTEREST: Indicates an interest in European geographical locations. This condition is triggered by the word "Europe", the presence of a European country name or a university name from the MastersPortal database in the query.
- TITLE MATCH: Indicates a specific interest in the programme to which the user is referred by Google. This condition is triggered by the user's entire query string being present in the title of the programme referred to, or all terms in the programme title being present in the user's search query.

NO INTEREST: Indicates none of the above conditions are met.

5.2.3 Academic Discipline

Each Master's programme listed on the MastersPortal.eu website has at least one academic discipline assigned to it. MastersPortal's database contains around one hundred academic disciplines in a two level hierarchy, an overview is provided in *Appendix K*. This is the same hierarchy of academic disciplines as used by the baseline recommendation evaluated during the first experiment and introduced in *Chapter 2.2.1*.

Academic Discipline	Programmes	% of Total
Law	1.843	9%
Environmental Sciences	3.034	15%
Engineering & Technology	2.328	12%
Business & Economics	1.666	8%
Humanities & Art	1.465	7%
Life Sciences, Medicine & Health	1.843	9%
Natural Sciences	3.423	17%
Applied Sciences, Professions & Arts	3.298	17%
Social Sciences	864	4%

Table 6: Top-level academic disciplines and their number of programmes in the MastersPortal database

There are nine top-level disciplines. For the purpose of the contextual analysis the second level of disciplines is discarded. All programmes are grouped under the nine top-level disciplines listed in *Table 6* above.

A programme can be assigned to multiple disciplines. The total number of programmes listed in the table above is thus higher than the total number of programmes in the database during the experiment. There is no way to determine the leading discipline for a programme. I have thus opted to count programmes with multiple top-level disciplines multiple times, once for each top-level discipline attached to the programme.

5.2.4 Screen Resolution

The screen resolution context provides information on what visitors see on the MastersPortal.eu website before they interact with it. This is especially useful because it tells us what happens if important content "flows" outside of the initially visible part of the website.

In order to make the diverse set of potential screen resolutions more manageable, the screen resolution context is divided up into five, mutually exclusive, categories. The categories are listed below in *Table 7*.

Category	W x H (pixels)
Mobile	< 600 x 350
Tiny	> Mobile < 800 x 450
Small	> Tiny < 1000 x 550
Medium	> Small < 1150 x 750
Large	≥ 1150 x 750

Table 7: Overview of the categories within the "Screen Resolution" context

The categories were constructed based on how many recommendations a typical visitor would be able to see without scrolling their browser window. By separating visitors based on these categories we can estimate how "overflowing" of the "Related Programmes" box affects the performance of the recommendations presented.

Visitors with a "Tiny" screen resolution see at most one recommendation. Visitors with a "Small" screen resolution see at most four recommendations. Visitors with a "Medium" resolution see between four and all recommendations. Visitors with a "Large" resolution see all recommendations. The "Mobile" category, capturing visitors using mobile phones, is not taken into account during the analysis. Mobile browsers represent a small portion of visitors *and* have greatly altered dynamics when compared to regular visitors. They are thus excluded from further analysis.
The screenshot provided in *Appendix C2* gives a visual overview of how much of the programme recommender is visible to each of the screen resolution categories.

Screen resolution cannot be detected server-side. Some client-side code is required to capture the information. The detected resolution is send to the server as separate request *after* the page has loaded and all other information is stored. As a result, some visitors did not properly communicate their screen resolution.
The screen resolution values noted in *Table 7* are based upon the visitor's "viewport" size. As such, they do not take into account the interface placed around the MastersPortal.eu website by the visitor's browser and their operating system.

5.3 Results

Now that the different contexts are defined, a post-analysis is executed on the data gathered during the first experiment. As this analysis is based upon the data gathered during the first experiment, not all five contexts previously discussed could be utilised.

The data required for the "Screen Resolution" context was not stored during the first experiment. This context is thus not taken into account in this chapter; it will be considered in *Chapter 7*.

5.3.1 Overall Bounce-Rate

The graph in *Figure 16* provides an overview of the bounce-rates per experiment group as determined during the first experiment analysis in *Chapter 3*. A clear distinction between the three homogeneous groups identified in the experiment is visible.

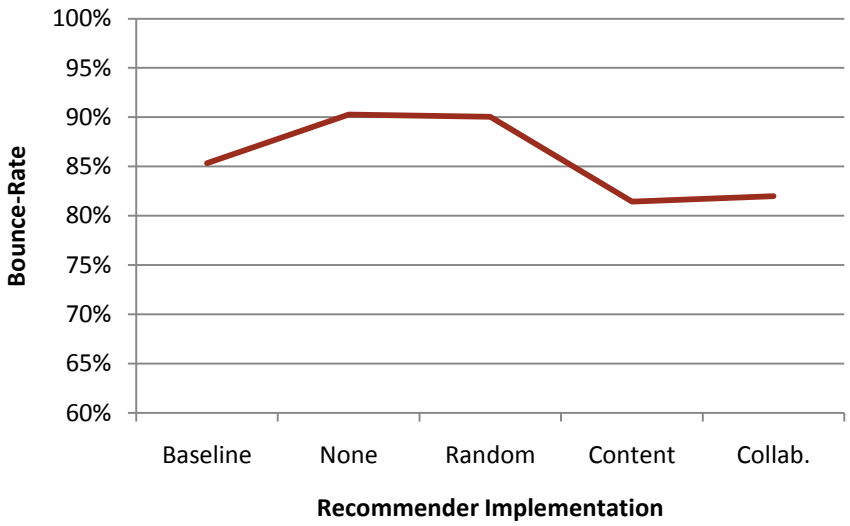


Figure 16: Overall bounce-rate during the first experiment for each recommender approach

All further graphs in this chapter use a similar axis configuration. This provides for an optimal comparability of the effects of the contextual factors against the overall results displayed in the graph above.

Apart from a visual overview of each contextual factor, a table with the results of the classification of visitors into contextual categories for each factor is also included. In these tables the number of sessions within each contextual category is provided. The number of sessions can be used to judge the stability of the category's preference.

There are two reasons why a classification based on contextual factors is undecided: Firstly, it can be caused by insufficient data; a low number of sessions. Secondly, there might be an inherent indecision within the classification. During the discussion in the next sections, the number of sessions within each category is evaluated to provide an insight into which of these two potential causes is most likely responsible for the undecided preference.

5.3.2 Geographical Origin

The first context analysed is the geographical origin of the visitor. For this analysis, the context was split up into a continental and a country specific part. A distinction is made between visitors from six continents, Antarctica excluded, and the top nine countries in number of visitors over the course of the experiment. The tenth country, The Netherlands, is excluded to prevent any undue influences

resulting from the presence of the MastersPortal's offices in the country.
The number of sessions for each of the regions within the context is displayed in *Table 8* below. Note that the number of sessions for each continent includes the number of sessions for the top nine countries within the continent. For example, the number of sessions for all countries in Europe, except Germany and the United Kingdom, is 16.009.

Region	Sessions
Africa	**4.138**
Nigeria	818
Asia	**14.903**
China	606
India	4.756
Iran	595
Pakistan	1.237
Turkey	1.029
Europe	**28.411**
Germany	2.419
United Kingdom	9.983
North America	**9.233**
United States	6.431
Oceania	**869**
South America	**1.062**

Table 8: Number of sessions recorded per region during the first experiment

The results of the analysis based upon a visitor classification over the six continents are displayed in *Figure 17* below. Considering this graph, both Africa and Oceania show a clear preference for collaborative recommender. The other continents seem to be largely in line with the overall result sketched in *Figure 16*.

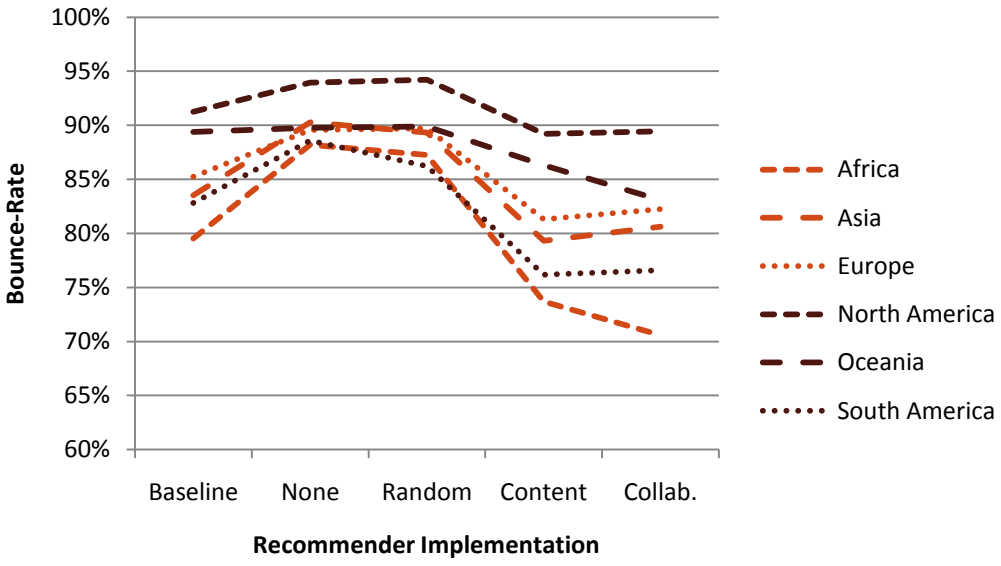


Figure 17: Bounce-rates for the first experiment when visitors are classified by continent

In the case of Africa the collaborative recommender performs significantly better than the content-based recommender. Oceania does not provide a statistically significant difference. At a reduced confidence interval there still are too few participants to provide a significant result. Looking at the

performance for Oceania in more detail, we see it deviates clearly from the pattern established in *Figure 16*, providing another indication that the number of participants is simply too small.

Looking more closely at the performance of the content-based and collaborative recommenders within Europe and Asia we see that both have a slight preference for the content-based approach. The content-based recommender performs significantly better for both continents, albeit just barely within the confidence interval.
The relative performance of each continent as compared to the others does differ quite significantly, but the internal differences inside each continent are quite similar for all four of the remaining continents.

Region	Content vs. Collaborative	Sessions
World	Undecided	-
Africa	Collaborative	4.138
Asia	Content	14.903
Europe	Content	28.411
North America	Undecided	9.233
Oceania	Undecided	869
South America	Undecided	1.062

Table 9: Recommender preference of visitors, classified by continent, during the first experiment

The results from the continental contextual factor are summarised in *Table 9*. Two continents prefer the content-based recommender. One prefers the collaborative recommender and three are undecided.
From the number of sessions per continent listed in the table we conclude that the "undecided" result of North America is most likely stable. This is caused by visitors from North America being similarly distributed to the overall population; their preference closely follows that of the entire visitor population. Alternatively, and less likely, it points to the fact that visitors from North America have no explicit preference for either recommender approach.
The number of sessions for the other two continents is much lower; pointing to the conclusion that for these two continents the number of participants might simply be too few to reach a decision.

The graph in *Figure 18* below shows the performance of the top nine countries in more detail. This graph shows similar, but more distinctive, patterns amongst the countries. The spread of recommender performance *within* the countries is much greater. The overall spread *between* the countries is nonetheless quite similar to that of the continents in *Figure 17*. The main pattern present in *Figure 16* is clearly visible for most countries.

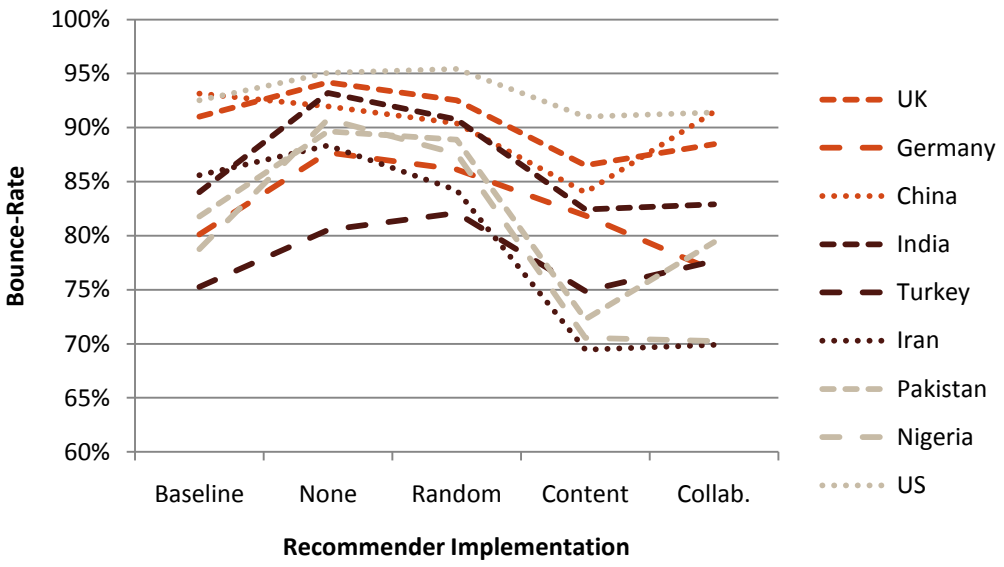


Figure 18: Bounce-rates for the first experiment when visitors are classified by country

In order to gain a better insight in the differences between the nine countries in the graph above, the graphs is split up into three graphs containing subsets of countries. The resulting graphs are provided in *Appendix I1*, their interpretation and results are discussed below.

China and Pakistan show a significant performance improvement of the content-based recommender over the collaborative recommender. Despite of the low number of sessions China has, the content-based recommender performs significantly better than the other recommendation approaches. Interestingly, all other recommendation approaches performs statistically equal in the case of China. Roughly the same goes for Pakistan: The content-based recommender performs significantly better. The collaborative recommender performs equal to the baseline recommender. In the case of Pakistan, the “none” and “random” control approaches perform significantly worse.

In Germany, the collaborative recommender performs significantly better than content-based approach, which is at the same level as the baseline recommender.
Why the collaborative recommender performs better in both Africa and Germany is not clear. There is a multitude of potential explanations to be found in the data gathered. It is out of scope for this thesis to try and confirm or falsify these explanations; it would require additional experimentation. As such, we will not attempt to provide any confirmed explanations as to why to behaviour of visitor categories is as observed.

Looking at Iran and Nigeria we see a pattern resembling the overall pattern from *Figure 16* considering the content-based and collaborative recommender approaches. Both countries do seem to make a strong distinction between the two automated approaches as opposed to the baseline recommender and the control groups.

The United States, the United Kingdom, India and Turkey show a very close resemblance to the overall pattern from *Figure 16***.** The United Kingdom and Turkey seem to prefer the content-based recommender over the collaborative recommender. Only the preference of the United Kingdom is significant though.

Region	Content vs. Collaborative	Sessions
World	**Undecided**	**-**
Africa	**Collaborative**	**4.138**
Nigeria	Undecided	818
Asia	**Content**	**14.903**
China	Content	606
India	Undecided	4.756
Iran	Undecided	595
Pakistan	Content	1.237
Turkey	Undecided	1.029
Europe	**Content**	**28.411**
Germany	Collaborative	2.419
United Kingdom	Content	9.983
North America	**Undecided**	**9.233**
United States	Undecided	6.431

Table 10: Recommender preference of visitors, classified by country, during the first experiment

The results for the country contextual factor are summarised in *Table 10* above. Germany seems to prefer the collaborative approach. Three other countries prefer the content-based approach and the remaining countries are undecided. It is difficult to determine which of the countries are undecided because of insufficient data and which are truly undecided. There does not appear to be a clear pattern.

From *Table 9* and *Table 10* we can clearly see that classifying visitors based on a geographical context allows for more statistical significance than can be achieved by looking at the combined results for all visitors.

Furthermore, it is striking to see how much higher the overall bounce-rate for North America is as opposed to the other continents. A similar pattern is discernable at the country level, where the United Stated, the United Kingdom, China and India have the highest bounce-rates. Although I do not want to draw any firm conclusions, it is interesting to note that from MastersPortal's experience we know that these four countries have a strong preference towards *not* studying in the Europe. Studying in the United States is very popular in India, China and the U.S. itself. Many students from the United Kingdom also prefer the U.K. and the U.S. over continental Europe. The higher bounce-rates for these regions could thus very well be explained by a lower interest in studying in Europe.

5.3.3 Google Query

The second context analysed is that of the query entered by the visitor on Google before being referred to the MastersPortal.eu website. As noted in the previous chapter, a simple textual analysis is performed on the query and based on the results of this analysis each visitor's sessions are classified into either the "*No interest*" or "*Interest in Master's Degree*" category.
In *Table 11* the number of sessions found to be in the different categories is listed. For the current analysis, each of the sub-categories of "*Interest in Master's Degree*" is setup to be mutually exclusive. So, the count for the "Europe" category contains *only* queries referencing to a general interest in European Masters. It specifically does *not* include queries that references to a country name, a university name or have a "title match".

Google Query	Sessions
Interest in Master's Degree	**36.471**
Europe	765
Country	2.962
University	1.007
Title Match	4.295
No Interest	**20.435**

Table 11: Number of sessions recorded per "Google query" during the first experiment

The graph in *Figure 19*, on the next page, shows the overall performance, from *Figure 16*, combined with the two main distinctions made based upon the visitors' Google queries. It shows that visitors with an interest in a Master's degree have a much lower bounce-rate than visitors without this interest. This confirms the simple textual analysis properly performs its task.
It is difficult to separate all interested visitors from the uninterested visitors. The current filter undoubtedly misses interested visitors. It is thus highly likely that uninterested visitors in reality have a higher bounce-rate.

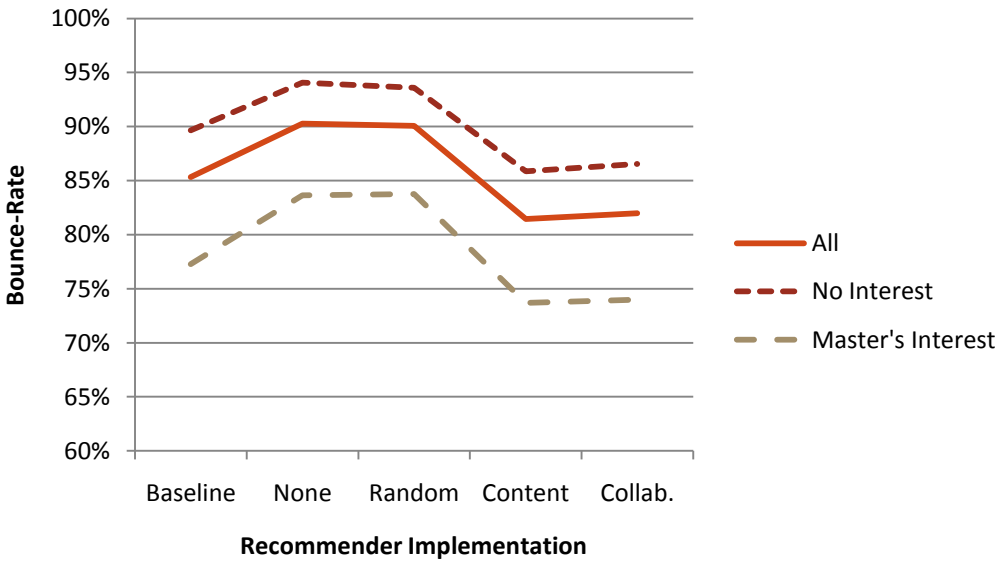


Figure 19: Bounce-rates for the first experiment when visitors are classified by academic interest

The graph in *Figure 20* provides an overview of the four sub-classifications made within the "Interest in Master's Degree" classification. A general observation becomes clear quickly: Visitors with a broad focus have a lower bounce-rate than visitors with a narrow focus.
Visitors with an interest in the broader geographical entities of "Europe" and the different countries have a significantly lower bounce-rate than visitors interested in a specific country.

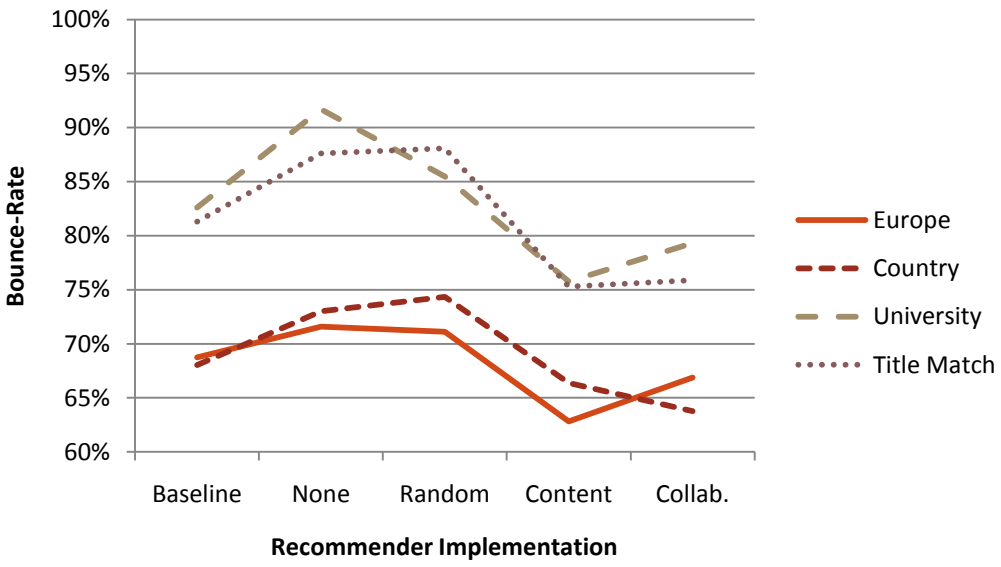


Figure 20: Bounce-rates for the first experiment when visitors are classified by specific academic interest

Visitors in the "Europe" contextual category have a wide focus and as such seem to prefer the content-based recommendation. Interestingly, they don't seem to care much for the collaborative recommendation, which scores effectively the same as the remaining three approaches. This category has too few participants to draw a statistically significant conclusion.

Visitors with a country-level interest seem to prefer the collaborative recommendation over the content-based approach. A potential explanation for this is that collaborative approach has a slight country-level bias. This bias is introduced by visitors browsing through the MastersPortal.eu website. It is possible for visitors to go through lists of programmes on a country and university level. Visitors doing this create a bias within the current collaborative recommender. The content-based and collaborative recommendations perform homogenously, again due to the relatively small number of participants.

Visitors looking specifically for a university overall seem to be most influenced by the presence of a recommendation. This is made clear by their very high bounce-rate in the situation without a recommendation. For these visitors the content-based recommendation scores significantly better than the collaborative recommendation.
A qualitative review of the queries indicating interest in a university shows that most of these queries consist of a rather generic interest in *any* academic programme at said university. It seems that the content-based recommender is able to help these visitors find exactly the programme they're looking for better than the collaborative recommender can.
This leads to the *very* cautious explanation that the content-based recommender might have a slight university bias. Since most universities have a central marketing department producing their programme descriptions, the overlap in writing-style could influence the recommendations; favouring similarly worded texts.

Looking at the "title match" contextual category, we see it follows the overall pattern laid out in *Figure 16* closely. This is not surprising, as the title match category most likely includes the broadest group of visitors.
These visitors are all well served with a content-based recommendation: They are searching for

terms that match the title of the programme they are currently viewing. As such, a content-based recommendation using the programme's description provides them with relevant suggestions.

Google Query	Content vs. Collaborative	Sessions
Overall	**Undecided**	**-**
Interest in Master's Degree	**Undecided**	**36.471**
Europe	Undecided	765
Country	Undecided	2.962
University	Content	1.007
Title Match	Undecided	4.295
No Interest	**Undecided**	**20.435**

Table 12: Recommender preference of visitors, classified by "Google Query", during the first experiment

The results in *Table 12* provide a strong indication there is not enough data available to draw significant conclusions for the Google query contextual factor. Only the "interest in university" contextual category shows a significant preference for the content-based recommender. Judging by *Figure 20* I would have expected the same fate to befall the "interest in Europe" category. For the "interest in a country" category a preference towards the collaborative recommender was expected.

5.3.4 Academic Discipline

The third and final contextual factor taken into account during the post-analysis is the academic discipline of the Master's programme viewed by the visitor. In *Table 13* below the nine top-level academic disciplines present in the MastersPortal database are listed, including the number of sessions recorded for each discipline during the experiment.

Academic Discipline	Sessions	% of Total
Law	3.178	4%
Engineering & Technology	17.010	21%
Humanities & Art	6.964	8%
Life Sciences, Medicine & Health	7.378	9%
Natural Sciences	6.401	8%
Applied Sciences, Professions & Arts	7.925	10%
Social Sciences	12.913	16%
Business & Economics	15.333	19%
Environmental Sciences	5.261	6%

Table 13: Number of sessions recorded per academic discipline during the first experiment

Comparing the relative distribution of sessions amongst the disciplines against the relative number of programmes in each discipline, as seen in *Table 6*, we see that the "Engineering & Technology" discipline has more sessions than expected. Conversely, the "Law "disciplines receives significantly less sessions.
The other disciplines receive an equal number to what is to be expected with respect to their number of programmes. As it is not unlikely that some disciplines are more popular than others, these results do not provide any cause for further investigation.

Figure 21 below illustrates the performance difference between the nine top-level disciplines. At first glance it appears the differences are less outspoken than for previous two contexts.
The performance of each of the disciplines is relatively close to the overall performance, leading to a more condensed graph. As this does not improve the readability of the graph, it is split up into three

graphs containing smaller subsets of disciplines. The resulting figures are provided in *Appendix I2*, their interpretation and results are discussed below.

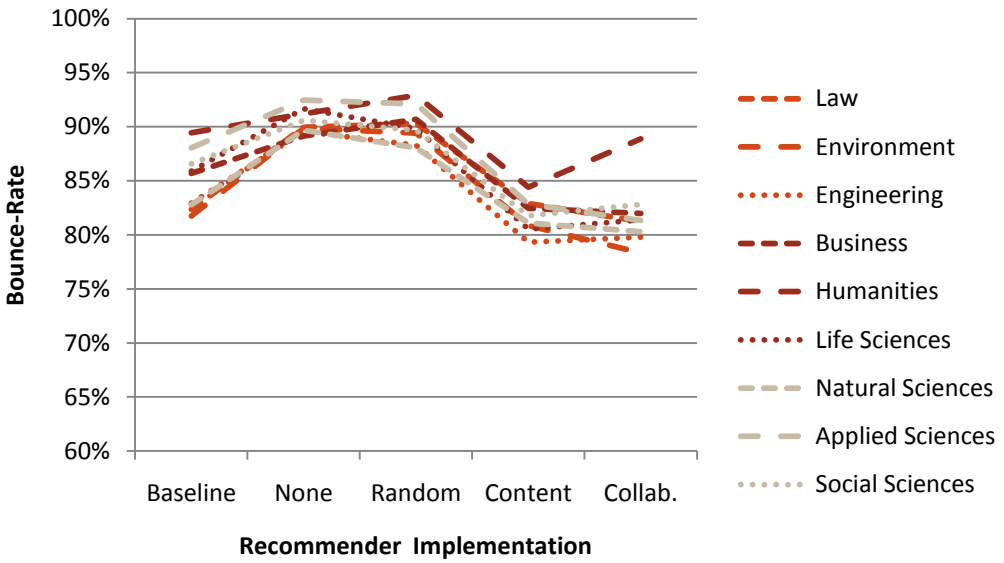


Figure 21: Bounce-rates for the first experiment when visitors are classified by academic discipline

The "Engineering & Technology" discipline most closely follows the overall pattern established in *Figure 16*. The "Life Sciences" and "Social Sciences" disciplines stay close to the overall pattern, but show a slightly more outspoken preference towards both the content-based and collaborative recommenders. For all three disciplines goes that the difference between the content-based and collaborative recommender is statistically insignificant. The "Social Sciences" discipline is a border-line case.

The "Environmental Sciences", "Applied Sciences" and "Law" disciplines all seem to prefer the collaborative recommender. Testing for statistical significance indicates the "Environmental Sciences" and "Applied Sciences" disciplines both provide a significant difference. The result for the "Law" discipline does not indicate significance. This is most likely caused by the fact that it is the smallest discipline within the MastersPortal database, both in number of sessions and number of programmes.
An interesting side note: Both the "Environmental Sciences" and "Applied" disciplines appreciate the baseline recommender just as much as the content-based recommender. These are the only two visitor categories in this thesis where this behaviour is observed.

The "Humanities & Arts", "Business & Economics" and "Natural Sciences" disciplines behave more erratic than the previous six; they provide no statistically significant results. Within the "Humanities & Arts" discipline, the collaborative recommender scores particularly bad. Furthermore, the "Business & Economics" discipline makes no distinction between the content-based and collaborative recommenders. It does show an interesting dislike for the random control group. The "Natural Sciences" discipline does not appear to make a distinction between either the content-based or collaborative recommenders either.

Academic Discipline	Content vs. Collaborative	Sessions
All Disciplines	Undecided	-
Law	Undecided	3.178
Engineering & Technology	Content	17.010
Humanities & Art	Content	6.964
Life Sciences, Medicine & Health	Content	7.378
Natural Sciences	Undecided	6.401
Applied Sciences, Professions & Arts	Collaborative	7.925
Social Sciences	Content	12.913
Business & Economics	Undecided	15.333
Environmental Sciences	Collaborative	5.261

Table 14: Recommender preference of visitors, classified by academic discipline, during the first experiment

An overview of the results for the "Academic discipline" context is provided in *Table 14*. We see two discipline categories favouring the collaborative recommender and four categories favouring the content-based recommender. The remaining three are undecided.

Looking at the results in *Table 14* there appears to be a less outspoken relation between the number of sessions in each contextual category and its ability to reach a stable decision; many of the larger visitor categories based on academic disciplines remain undecided.
Just as with the geographic origin contextual factor, this most likely points to the fact that within these disciplines participant distribution is so alike to overall distribution that the results follow those of the overall distribution closely. Again less likely is the conclusion that in fact visitors in these categories have no preference towards either of the recommender approaches.

5.4 Conclusions

The foremost conclusion drawn from the analysis on contextual factors is that more statistically significant differences arise when visitors are classified along the lines of a certain context. Hence, finding good contextual factors and proper classifications within these factors can improve the performance of a recommender system. This improved system will be able to better select a recommendation approach based upon preferences of similar visitors.

Contextual factors need to be chosen very carefully. From the results of the analysis presented in the previous chapter two important factors emerge.
Firstly, the contextual factor should allow for sufficient data to be gathered. Without sufficient data, no statistically significant conclusions can be drawn, even though the context provides potentially interesting results. This is the case for the "Google query" context.
Secondly, classifications within the contextual factor need to be specific. A context with several broad categories does not provide added statistical significance to the system. This is especially clear in the "academic discipline" context. Here we see various large disciplines that do not provide a statistically significant preference for either the content-based or collaborative recommenders.
It is of course possible that these disciplines do not have a preference, but it is more likely that visitors within these large disciplines are so heterogeneously distributed that the discipline itself does not differ enough from the average visitor. As the average visitors has no significant preference, the disciplines has not got one either.
The power of system utilising the contextual factors thus lies in its ability to establish sufficiently large homogeneous groups of visitors from the overall population.

The results of the analysis executed in this chapter also have an impact on the conclusions drawn for the first experiment in *Chapter 4.6*.
For each of the contextual factors analysed, either the content-based or collaborative recommender approach performed equally or better than any of the other approaches. This strengthens the conclusions drawn in *Chapter 4.6*: If we need to select one of the two recommender approaches, either the content-based approach or the collaborative approach perform equally well on the MastersPortal.eu website based on the results of the first experiment.
Concerning the collaborative recommender, indications of bias are apparent in the results presented in this chapter. This bias is introduced by not properly filtering the harvested behavioural data. If the collaborative recommender is to be put to practical use, we will need to carefully guard against introducing biases through the behavioural data.

Overall, the results of the analysis show that there is more merit to the collaborative recommender than can be assumed through the conclusions of *Chapter 3*. In all contextual factors investigated, there are categories of visitors in which the collaborative recommender approach performs better than the content-based approach. During the discussion of the contextual factors in the previous section, potential explanations for some of these differences were provided.
For the geographic origin contextual factor it appears as if the collaborative recommender is slightly biased towards certain regions; potentially providing visitors from these regions with a more relevant recommendation. This could be caused by visitors from these regions, on average, viewing more programmes than visitors from the rest of the world.
Within the Google query contextual factor we see that visitors with a strong preference towards the specific academic contents of a Master's programme, a "title match", prefer the content-based

recommender approach. Visitors with a more general interest are undecided.
Verifying these explanations is difficult, if not impossible. Although statistical differences are clear, a rational explanation can only be provided in some cases; in those cases there is no certainty the explanation is valid. It would therefore seem unwise to attempt to draw any generalised conclusions without constructing and executing further experiments to explicitly test these explanations.

There is thus no clear decision rule that can be applied to select between the content-based and collaborative recommender. It appears this determination can only be made through the analysis of visitor behaviour, either real-time or through previously recorded data.
A future implementation aiming to use both types of recommendations effectively will thus need to be, up to a certain degree, aware of these contextual factors. Preferably it should be able to tune its recommendations in real-time based upon the relative performance within the different contextual factors.

The results of the contextual factors analysis are verified through a second experiment whose results are presented in the next chapter. The remainder of this thesis subsequently focuses on the idea of using contextual factors to build an adaptive recommender system. The detailing of this idea can be found in *Chapter 7*.

6. Contextual Factors Experiment

The main conclusion drawn from the analysis of the contextual factors in the previous chapter is that there are clear differences in performance between the content-based and collaborative recommenders when visitors are classified based on certain contextual factors.
The results of the previous analysis, which is executed "ex post" based on the results of the first experiment, requires further verification. This chapter describes a second experiment that was executed to provide verification of the conclusions stemming from the previous chapter.

The second experiment focuses on a single contextual factor in order to keep the complexity within the scope of this thesis. The contextual factor taken into account in this experiment is the geographic origin of the visitor.
Apart from validating the results presented in the previous chapter, the second experiment serves as a data gathering mechanism for the final chapter of this thesis, an investigation into the potential for an *adaptive recommender* and *retrieval* system.

6.1 Experiment Design

The goal of the second experiment is to determine whether classifying visitors based on a certain context provides a different, more conclusive, indication of the relative performance of the two best recommender approaches from the first experiment.
The design of the experiment is closely related to the design of the first experiment as described in *Chapter 4.1*. As such only the differences between these two experiments are subsequently discussed in more detail.

During the discussions in the previous chapter three contextual factors were analysed; only the visitor's "geographical origin" contextual factor is involved in this experiment. This contextual factor was chosen because it is easily and reliably measurable and provided interesting results in the previous chapter.
To further bring the complexity of the second experiment down, only a subset of the regions discussed in *Chapter 5.2.1* are included in the experiment. Regions are selected based upon the number of sessions counted towards each region during the first experiment.
Only regions with over 2.000 sessions as listed in *Table 8* and *Table 10* were included. This number was chosen because it provides a clear division between small and large regions within the dataset. This selection is based on the expectation that for the second experiment the relative distribution of regions will be the same.

The second experiment considers two recommender approaches: The content-based recommender and the collaborative recommender. During the first experiment, these two treatment groups performed equally well.
No control groups are defined for the second experiment. We are only interested in the difference between the two treatments; there is thus no need for further baseline measurements. An introduction on the two treatment groups utilised is provided in *Chapter 3.3*.

The effect of each recommender is judged by looking at the bounce-rate of visitors from a specific country who, during their visit, are presented with the respective recommender. The bounce-rate of a group of visitors is the fraction of their *visits* that only view a single page. Note that each visitor, over a certain period of time, can have multiple *visits* to the MastersPortal.eu website.

The experiment again takes the form of a controlled experiment in which all factors, apart from the recommender approach, are kept unchanged throughout the experiment.

Participants are equally divided amongst the two experiment groups using a round-robin assignment based on their country of origin. The first participant from a certain country is randomly assigned to one of the two groups. The second participant is assigned to the opposite group; the third participant is again assigned to the opposite group and so on.
This round-robin approach assures an optimal distribution of the groups within all countries and as a result within all continents**.**

Intuitively, it would seem better to alternate a single participant between the two recommender approaches. Doing this would provide a better indication of the relative performance of each of the recommenders. The problem with this approach is that we are using the bounce-rate of the participants as our performance metric. By definition, participants bounce if they view only a single page during their visit. As a result, the participants we are interested in will always be part of a single group. It thus has no effect to alternate the recommendation group within each participant.

6.1.1 Conceptual Model

The overall design of the second experiment as described above is transferred into a conceptual model, which is provided below in *Figure 22*. The model in this figure provides a high level overview of the flow of the second experiment.

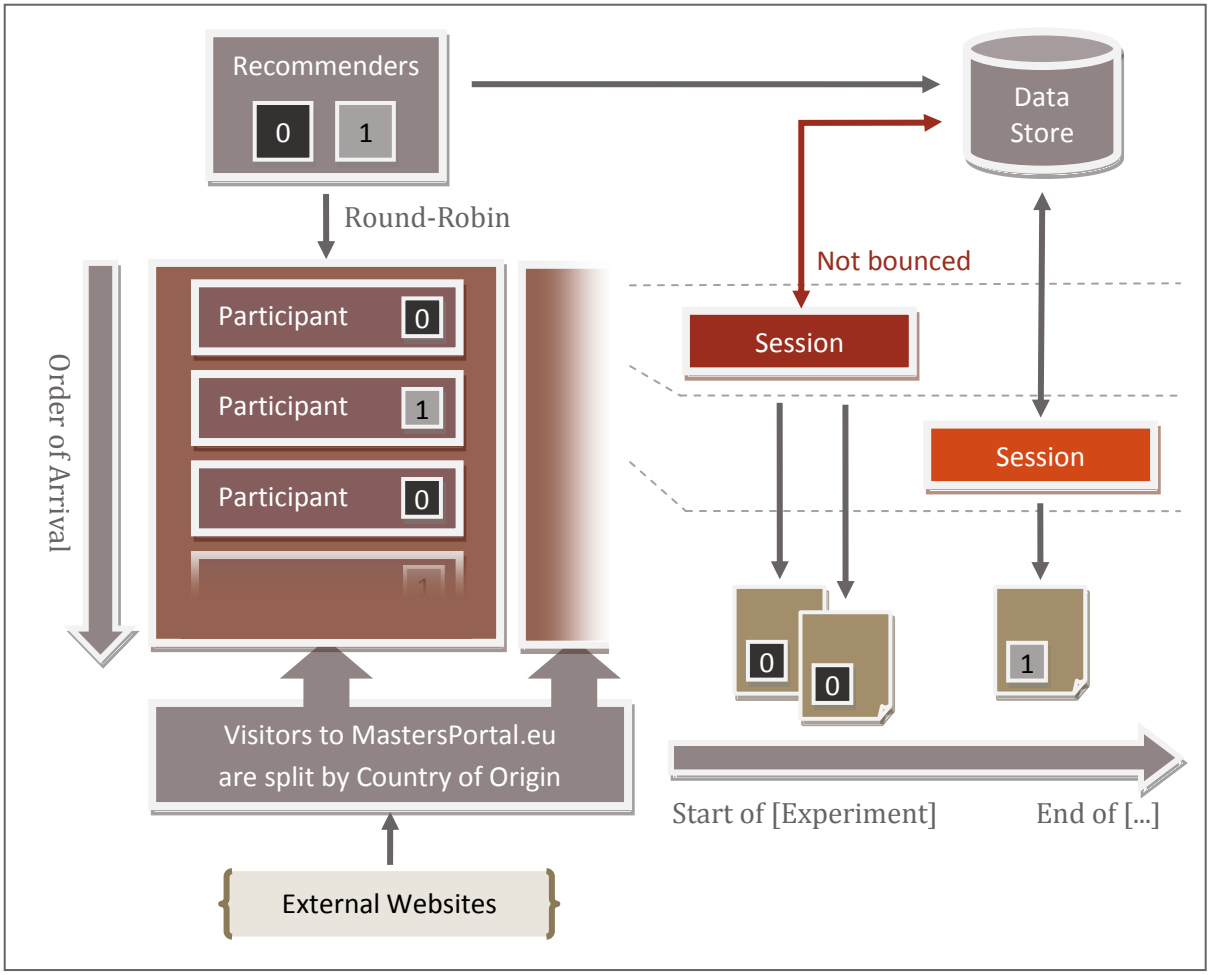


Figure 22: Conceptual model of the second experiment

6.2 Hypotheses

Based on the experiment design discussed in the previous chapter, three hypotheses are formulated for the second experiment[7].

Hypothesis 4: Visitors from Africa and from Germany prefer a collaborative recommendation over a content-based recommendation.

Hypothesis 5: Visitors from Asia and Europe and visitors from the United Kingdom prefer a content-based recommendation over a collaborative recommendation.

Hypothesis 6: Visitors from North America and visitors from India and the United States prefer either the content-based recommendation or the collaborative recommendation.

The hypotheses discussed in this chapter are more practical in their nature. This is a direct result of the goal of this experiment: To confirm the influences of a visitor's geographic origin as uncovered in the previous analysis.
Details on the exact statistical procedure utilised are again provided in *Chapter 2.5* of the background discussion for this thesis.

6.2.1 Hypothesis 4

The intention of the fourth hypothesis is to show that there are geographic regions where the collaborative recommendation outperforms the content-based recommendation. To properly test this hypothesis it is split up into two sub-hypotheses.

Hypothesis 4.1: Visitors from Africa prefer a collaborative recommendation over a content-based recommendation.

H0: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Collaborative"* is lower;

Hypothesis 4.2: Visitors from Germany prefer a collaborative recommendation over a content-based recommendation.

H0: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Collaborative"* is lower;

6.2.2 Hypothesis 5

The fifth hypothesis consists of the regions that showed a strong preference towards the content-based recommendation. The goal of this hypothesis is to show there are regions where the content-based recommendation outperforms the collaborative recommendation. In order to properly test this hypothesis it is split up into three sub-hypotheses.

[7] To avoid confusion the hypotheses are numbered consecutively to the hypotheses discussed in *Chapter 4.2*.

Hypothesis 5.1: Visitors from Asia prefer a content-based recommendation over a collaborative recommendation.

H0: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Content"* is lower;

Hypothesis 5.2: Visitors from Europe prefer a content-based recommendation over a collaborative recommendation.

H0: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Content"* is lower;

Hypothesis 5.3: Visitors from the United Kingdom prefer a content-based recommendation a collaborative recommendation.

H0: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Content"* is lower;

6.2.3 Hypothesis 6

The sixth and final hypothesis contains the regions which were undecided in the analysis on contextual factors. The goal of this hypothesis is to show that there are regions in which both recommendation approaches perform equally well. The sixth hypothesis is split up into three sub-hypotheses.
Due to the way hypothesis tests are formulated, for *Hypothesis 6* to confirm to our expectation it should be *rejected*. This as opposed to the other hypothesis tests in this chapter; there the expected result is to accept the hypothesis.

Hypothesis 6.1: Visitors from North America prefer either the content-based recommendation or the collaborative recommendation.

H0: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are not equal;

Hypothesis 6.2: Visitors from India prefer either the content-based recommendation or the collaborative recommendation.

H0: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are not equal;

Hypothesis 6.3: Visitors from the United States prefer either the content-based recommendation or the collaborative recommendation.

H0: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are equal;
H1: The bounce-rate for *Treatment "Content"* and *Treatment "Collaborative"* are not equal;

6.3 Experiment Setup

This chapter focuses on the practical implementation of the second experiment. It details the experiment's execution and highlights any further assumptions and limitations introduced based upon the implementation.

The second experiment focuses purely on the content-based recommendation and the collaborative recommendation. A detailed description of their implementation is provided in *Chapter 4.3.1*. The three other groups from the first experiment are completely discarded. No further changes were made to the recommenders as part of the second experiment. All other assumptions made during the first experiment and the conditions under which it was executed are kept exactly the same for the second experiment.

6.3.1 Execution of the Experiment

All visitors to the English-language version of the MastersPortal.eu website again participate in the experiment. The recommendation itself is presented on all pages that provide detailed information on a single Master's programme. All *visits* referred by Google to one of these pages are taken into account. All other visits are ignored.

The experiment runs for two weeks[8]. Each new participant is entered into one of the two experimentation groups using a round-robin assignment approach based on their country of origin. Once assigned, the participant remains in this group for the duration of the experiment.

6.3.2 Future Changes in Experiment Setup

Considering the simplified setup of the second experiment, the extensive filtering applied during the first experiment is most likely not entirely necessary.
For the second experiment it was decided to be better to re-utilise the existing design, than to re-engineering the entire process. The existing design had already proven itself and was thus a more logical choice given the time constraints.

For future experiments on the MastersPortal.eu website a simpler setup is possible. An important element that can be omitted for many experiments is the identification of visitors through cookies. Leaving this element out would simplify the experiment design and lead to much less loss of participants due to (potential) cookie issues.
Tracking-cookies are currently used to strictly enforce the controlled nature of the experiment. For many future experimental setups, such a strict enforcement of the experiment conditions is most likely not necessary. Adding to this, the reliance on a tracking-cookie can by itself be a limiting factor for the experiment. As not all visitors accept cookies, a portion of the visitor demographic is completely excluded from all experiments. It might well be that this part of the demographic behaves different from the part that does accept cookies.

[8] Due to time constraints a further extension of the experimentation interval is not possible. Results from the first experiment indicate the interval is sufficient to provide statistically significant conclusions on the larger regional contexts as discussed in *Chapter 5*.

6.4 Results

The experiment ran from March 17th of 2010 up to and including March 30th of 2010 on www.mastersportal.eu, providing two full weeks of data. The final results of the experiment were generated through an offline analysis.
As the analysis discussed in this chapter closely follows the procedure described in *Chapter 4.4* the result overview in the following chapter is kept intentionally brief.

6.4.1 Results Overview

The daily number of participants created during the second experiment is displayed in *Figure 23* below. Due to time constraints on the execution of the experiment no additional data beyond the two weeks displayed in the graph was gathered. As previously indicated in *Chapter 4.4.1*, the extension of the experimentation interval with a cool-down period has no effect on its results.

Filtering of invalid participants was again applied to remove as much noise as possible from the dataset. The same filtering procedure as applied during the first experiment was used. *Figure 10* in *Chapter 4.3.2* provides a schematic overview.
A grand total of 90.650 participants were recorded during over course of the experiment.
After filtering non-human visitors and participants who did not accept the experimentation cookie, 54.265 participants remained.
Of these participants, 46.792 saw only complete recommendations. The other participants either saw one or more incomplete recommendations, or viewed a page where no recommendation was present at all. As new programmes are added to the MastersPortal database continuously, the recommenders were not able to provide a recommendation for every programme.

The graph in *Figure 23* also provides an overview of the number of participants after all filtering was applied. A detailed description of the filtering process is presented in *Appendix F2*.

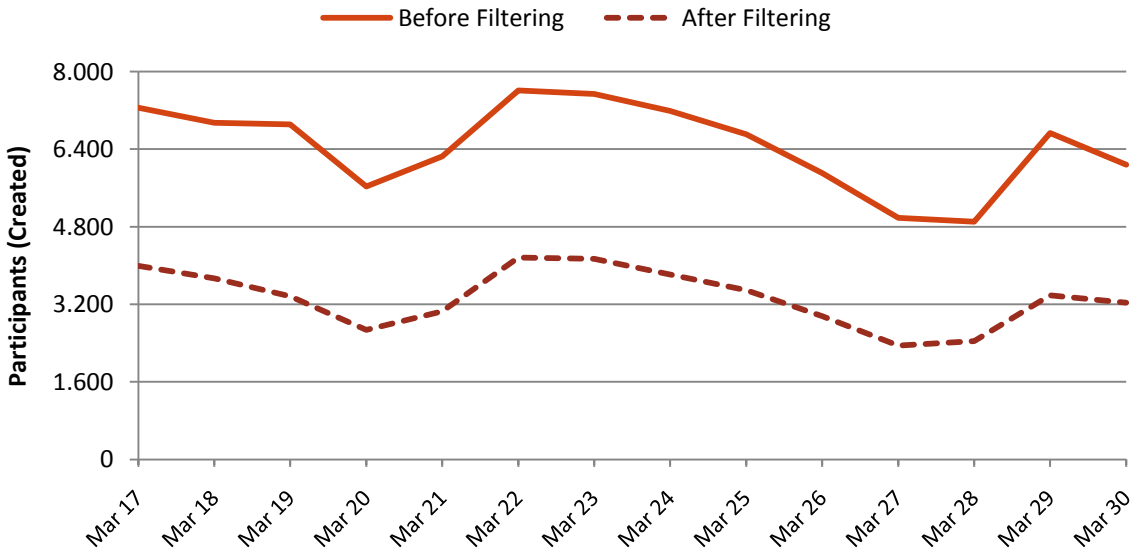


Figure 23: Daily number of participants created during the second experiment, before and after filtering

As we are only interested in visitors referred to the MastersPortal.eu website through Google, all sessions not started at Google were again excluded. This caused the number of sessions to drop from 51.490 to 39.411. The result of this final filtering step is displayed in *Figure 24*.

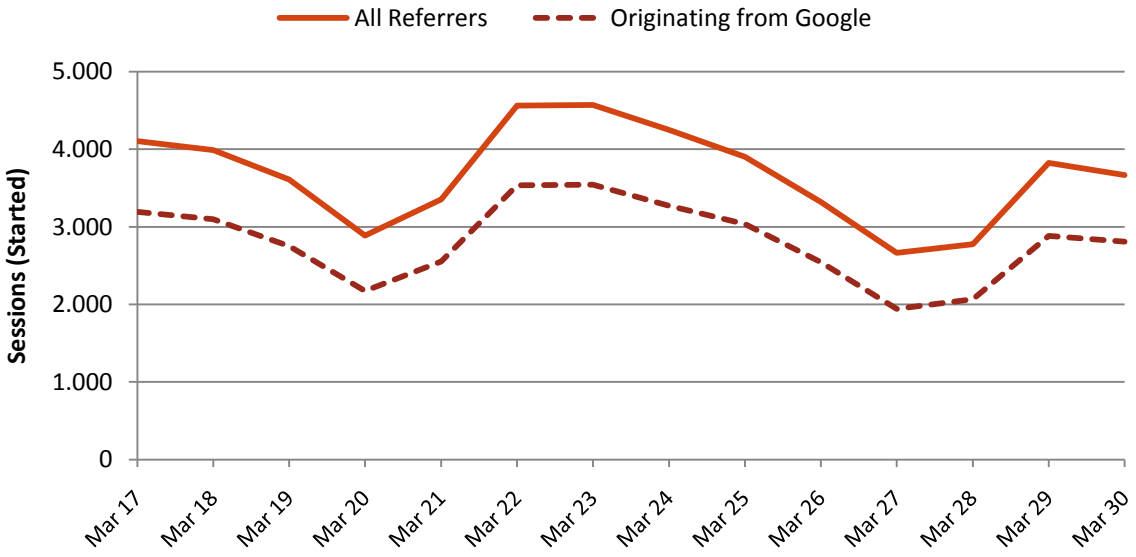


Figure 24: Daily number of sessions started, overall and originating from Google, during the second experiment

The final step before considering the hypotheses is analysing the distribution of participants amongst the two experiment groups. Participants should be distributed equally amongst these two groups. An overview of the grouping is presented in *Figure 25*.

Considering the unfiltered participants, $49{,}9\%$ was assigned to the "content" treatment group and $50{,}1\%$ was assigned to the "collaborative" treatment group.
The collaborative recommender is again somewhat more prone to providing incomplete recommendations. After all filtering is applied the distribution of participants between the "content" and "collaborative" treatments becomes slightly skewed: The "content" treatment receives $51{,}6\%$ of the participants and the "collaborative" treatment receives $48{,}4\%$ of the participants. This difference is in line with the difference observed after filtering the results of the first experiment.

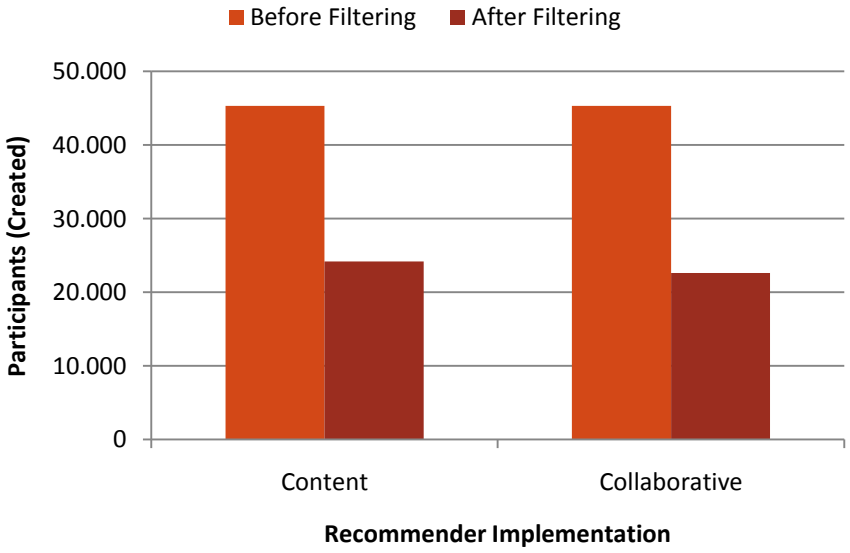


Figure 25: Total number of participants for each of the experiment groups during the second experiment

As discussed in *Chapter 4.4* the imbalance caused by the filtering does not influence the overall validity of the results, it merely complicates its analysis.

Finally, from *Figure 25* one other interesting observation can be made: The difference between the unfiltered and filtered number of participants is much greater than during the first experiment. This difference is caused by a single non-human visitor who managed to "create" over 20.000 invalid participants during the second experiment. By not accepting the experimentation cookie, each request send out by this "visitor" caused a new participant to be added to the experiment. More details are provided in *Appendix F2*.

6.4.2 Welch's T-Test

After completing final data selection, the bounce-rate for the two groups within the experiment was computed. In this chapter, the bounce-rates computed are interpreted using Welch's t-test procedure. The goal of this analysis is to show the preference of each of the regions towards either one of the two recommender approaches.
As the bounce-rate is a binomial variable in this experiment, it is approximated normally for the purpose of applying Welch's t-test. Statistical significance is set at 99% for all analyses in this chapter. The details of why this approach was chosen and the exact testing procedure are again outlined in *Chapter 2.5*.

Firstly, the results of the t-test executed on the data gathered during the second experiment are presented. Subsequently, the results from the post-analysis of the previous chapter are provided in a similar format. This is done to facilitate the comparison between the results. Note that the results for the contextual factors analysis were computed using a multiple comparison procedure, as opposed to the t-test procedure utilised for the second experiment.

Results of the Second Experiment

Table 15 below shows the number of sessions in both experiment groups for the regions taken into considering in the hypotheses presented at the start of this chapter. In *Table 16* the bounce-rates for the two experiment groups and the difference between the groups are noted.

Region	[Collaborative]	[Content] Sessions
Africa	**1.472**	**1.560**
Asia	**4.947**	**5.205**
India	1.789	1.905
Europe	**9.943**	**9.179**
Germany	812	845
United Kingdom	3.296	3.783
North America	**2.929**	**3.177**
United States	2.070	2.254

Table 15: Number of sessions per group during the second experiment

Looking at the results in *Table 16*, we see that both Africa and Germany still have a strong preference towards the collaborative recommender. The preference from within Africa is somewhat less than it was during the contextual factors analysis. Europe as a whole still prefers the content-based recommender.

Interestingly, the preference of visitors from Asia has all but reached a tie. The previously insignificant results for North America and the United States now show a significant preference towards the content-based recommendation.

Region	[Collaborative]	[Content] Bounce-rate	Delta[9]
World	**82,21%**	**81,58%**	**0,63%**
Africa *	**74,59%**	**76,99%**	**-2,39%**
Asia	**80,01%**	**79,90%**	**0,10%**
India	81,10%	81,05%	0,05%
Europe *	**82,39%**	**81,57%**	**0,82%**
Germany *	74,51%	79,41%	-4,90%
United Kingdom *	88,32%	86,52%	1,80%
North America *	**89,07%**	**87,32%**	**1,76%**
United States *	90,58%	89,26%	1,32%

Table 16: Bounce-rate per region during the second experiment; * indicates statistically significant delta

Results of the First Experiment

The results of the geographic origin contextual factor as presented in *Chapter 5.2.1* are summarised below. They are provided in a format similar to the results presented for the second experiment to facilitate their comparison.

Region	[Collaborative]	[Content] Sessions
Africa	**768**	**893**
Asia	**2.827**	**2.968**
India	906	973
Europe	**5.195**	**5.699**
Germany	491	463
United Kingdom	1.744	2.023
North America	**1.754**	**1.842**
United States	1.255	1.259

Table 17: Number of sessions per group during the first experiment

Table 17 above provides an overview of the number of sessions in each experiment group for the relevant geographic entities. In *Table 18* the bounce-rates for each of the two experiment groups and the difference between the two are noted.

Region	[Collaborative]	[Content] Bounce-rate	Delta[9]
World	**82,07%**	**81,45%**	**0,62%**
Africa *	**70,57%**	**73,68%**	**-3,11%**
Asia *	**80,62%**	**79,31%**	**1,30%**
India	82,89%	82,43%	0,47%
Europe *	**82,23%**	**81,29%**	**0,94%**
Germany *	76,78%	81,86%	-5,08%
United Kingdom *	88,47%	86,51%	1,97%
North America	**89,45%**	**89,20%**	**0,26%**
United States	91,39%	91,02%	0,37%

Table 18: Bounce-rate per group during the first experiment; * indicates statistically significant delta

[9] Values in the table are rounded to two decimals. As a result, the delta noted can differ slightly from the actual difference of the two bounce-rates mentioned.

Comparing *Table 16* against *Table 18* we see that the “World” bounce-rates have remained stable, both in absolute values and when looking at the delta value. Within the regional contexts significant shifts are visible though. Africa for example has a much higher average bounce-rate, whereas both Germany and India have a much lower average bounce-rate.

6.5 Validation

As a result of the simplified setup of the second experiment and the fact that virtually its entire implementation and analysis procedure is copied from the first experiment *as-is*, less rigorous validation is required. For the second experiment only a single validation against Webalizer is provided. The complete results of this validation are available in *Appendix G2*.

Looking at the results provided by Webalizer we see no discrepancies between the two sets of visitor statistics. No strange spikes are present and both the results generated by the experiment and those generated by Webalizer are closely correlated.

To further judge the validity of the results gathered from the second experiment, the values of two key metrics from the first experiment are compared to those of the second experiment. Combined with the previous Webalizer comparison this provides enough evidence on the validity of the second experiment.

6.5.1 Comparison with First Experiment

These metrics used in the comparison are computed on the unfiltered data. These metrics aim to further show the second experiment functions properly. They provide no indication on the proper functioning of the filtering applied to the results prior to their analysis.

Number of Recommendations per Session

The number of recommendations per sessions indicates how many Master's programmes are viewed during each session. If all participants bounce, this metric is exactly 1.

The average number of recommendations per session during the second experiment is slightly higher than during the first experiment. This is most likely a result of the nature of the second experiment. Considering that the second experiment utilises the two best performing recommenders from the first experiment it is likely to convince more participants to stay on the MastersPortal.eu website. Participants who stay on the MastersPortal.eu website receive further recommendations; this causes an increase in the average number of recommendations per participant.

During the first experiment, each participant received $1{,}32$ recommendation; during the second experiment each participant received $1{,}37$ recommendation.

Number of Sessions per Participant

The number of sessions per participant indicates how many visits the average participant pays to the MastersPortal.eu website. Just as previously, a visit is defined as a set of requests with at most sixty minutes in between them. If all participants visit only once, the metric is exactly 1.

The number of sessions per participant is nearly the same for both experiments. This provides a cautious indication to the fact that visitors do not return more often due to the perceived quality of the recommendation presented.

During the first experiment, each participant had on average $1{,}07$ sessions; during the second experiment, each participant had on average $1{,}05$ sessions.

6.6 Conclusions

By merely looking at the results of the analysis we can already conclude that the results of the second experiment indicate regional differences have an important impact on the relative performance of the two recommender approaches tested. This provides a confirmation of the conclusions drawn in *Chapter 5.4*.

Comparing the results from *Chapter 5* against those of the second experiment show several overlapping regional effects, but they also show marked differences between the two analyses. To better illustrate the differences and overlap, *Table 19* provides an overview of the results of all hypotheses posed at the start of this chapter.

Region	Result of Hypothesis Test		Expectation
Africa	**4.1**	**H^0 Rejected**	**Reject H^0**
Asia *	**5.1**	**H^0 Not Rejected**	**Reject H^0**
India	6.2	H^0 Not Rejected	Do not Reject H^0
Europe	**5.2**	**H^0 Rejected**	**Reject H^0**
Germany	4.2	H^0 Rejected	Reject H^0
United Kingdom	5.3	H^0 Rejected	Reject H^0
North America *	**6.1**	**H^0 Rejected**	**Do not reject H^0**
United States *	6.3	H^0 Rejected	Do not reject H^0

Table 19: Overview of results for hypotheses 4, 5 and 6; * indicates result not matching expectation

Based on *Table 19* the following conclusions are drawn with regard to the hypotheses:

Hypothesis 4: Visitors from Africa and from Germany prefer a collaborative recommendation over a content-based recommendation.

Visitors from both Africa and Germany show a statistically significant preference towards the collaborative recommender. As such the fourth hypothesis is accepted in full.

Hypothesis 5.1: Visitors from Asia neither prefer a content-based recommendation nor a collaborative recommendation.

Hypothesis 5.2 & 5.3: Visitors from Europe and visitors from the United Kingdom prefer a content-based recommendation over a collaborative recommendation.

The content-based preference shown by the Asian continent in *Chapter 5.3.2* has disappeared. Looking at the delta values listed in *Table 16* Asia now seems strongly undecided. The results for Europe and the United Kingdom are in accordance with our expectation.
This provides evidence of the fact that there are also regions which strongly favour the content-based recommendation over the collaborative recommendation. Although this might be considered implicitly from the overall results, it is important to show that the undecided overall result is not caused by a general indifference, but by different preferences throughout the contexts.

The fourth hypothesis furthermore points towards an inherent instability of the visitor recommender preference over time.

Hypothesis 6.1 & 6.3: Visitors from North America and the United States prefer a content-based recommendation over a collaborative recommendation.

Hypothesis 6.2: visitors from India neither prefer a content-based recommendation nor a collaborative recommendation.

The initially indifferent North American continent and the United States now show a strong preference for the content-based recommender. This is not in accordance with our expectations. The result for India is in accordance with our exception, but it needs to be noted here that the preference of India is only barely insignificant.
Overall I am thus inclined to state that *Hypothesis 6* needs to be fully rejected, even though statistically the result of India points to the opposite.

An important factor in the above differences likely lies in the fact that the MastersPortal.eu website has evolved considerably in the time that passed between the two experiments. The screenshots in *Figure 26* and *Figure 27* show the "highlights" bar of the MastersPortal.eu website, where the recommender experiment was located. Between January of 2010 and March of 2010 significant changes were made to the overall look and feel of this bar.
It is important to note that *during* the experiments, no significant layout changes were made to the MastersPortal.eu website.

Apart from changes to the layout of the MastersPortal.eu website, the large growth in visitor numbers the website has seen from January of 2010 to March of 2010 is most likely a factor too.
In January of 2010, the MastersPortal.eu website received around 700.000 monthly visits. In March of 2010, at which time the second experiment was executed, the website received nearly 900.000 monthly visits.

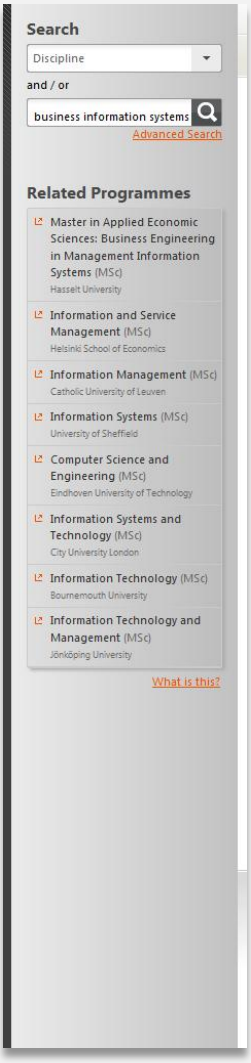


Figure 26: January 2010

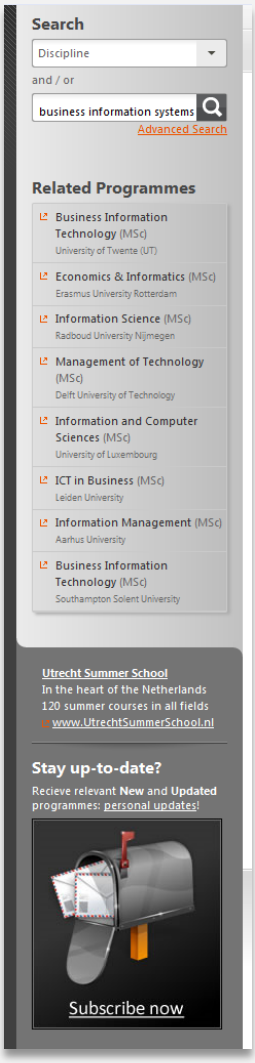


Figure 27: March 2010

The results of hypotheses provide a mixed picture. The rejection of *Hypothesis 4* is in line with my expectations. It provides strong evidence to the fact that certain geographical visitor categories do prefer the collaborative recommender over the content-based recommender. This again adds weight the notion that the collaborative recommender should definitely not be discarded.
The conclusions based on *Hypothesis 5* and *Hypothesis 6* are partially in line with my expectations but do show discrepancies. It is thus important to realise that the visitor preference is apparently not fixed over time. A one-time analysis will thus not suffice when the aim to provide each visitor during every visit with an optimal recommendation.

The results of the second experiment clearly point towards the important influence of the contextual factors. They also point towards the fact that the exact influence of the contexts may not be deducted logically and it may neither be fixed over time.
In order to implement a system utilising the results of the previous two chapters, an adaptive system is required. In the next chapter a final analysis is performed on the data gathered. The goal of this analysis is to judge whether in the future it will be possible to implement a system which adaptively uses the results of a contextual analysis to best tailor its recommendation to the preference of the visitor.
The analyses executed in the next chapter are based on an interpretation of the results gathered in this and the previous chapter. As such, the contents of the next chapter were not originally intended to be part of this thesis. The approach utilised is not as exhaustive as the one applied during in the current and previous chapters. The next chapter's intention is to provide future directions for research within MastersPortal and to satisfy, at least to some extent, the curiosity raised by the results of the experiments executed during this thesis.

7. Feasibility of an Adaptive Recommender System

The results of the second experiment show that contextual factors influence the relative performance of the content-based recommender and collaborative recommender. In order to optimally capitalise on these apparent differences in performance, a more dynamic solution to combining the recommenders is required.
This chapter aims to determine if it is feasible to use the concept of an *adaptive recommender system* within the MastersPortal.eu website. This chapter is included in the thesis based on the results gathered in *Chapters 5 and 6*. The goal of this chapter is *not* to implement an adaptive system, but to investigate whether it is possible to implement such a system around the contextual factors identified in the previous chapters. This chapter should be viewed as an addition to the thesis; not as one of its core components. Its main goal is to provide a future direction for research within the MastersPortal.

7.1 Introduction

At the core of the wish to implement an adaptive recommender system is the realisation that a static recommender system does not provide the best possible recommendations. From the previous chapters it is clear that generalised conclusions cannot be drawn based upon the results of a single statistical analysis. Although clear differences are apparent, it is often guesswork as to why a certain contextual category performs as it does. As such, setting up a static recommender system based upon the results presented in the previous chapters is not advisable.
Furthermore, the results of the second experiment indicate the effects of the contexts are not necessarily stable over time. What currently performs well might prove to be a bad choice in the future.

Based upon these considerations an adaptive recommender system seems to be a more suitable solution. As discussed in the previous chapter, the preference towards either one of the two recommender approaches becomes more outspoken when visitors are classified based upon contextual factors. If the recommender system takes these contextual visitor categories into account and adapts its recommendation accordingly it becomes *context-aware*.

As implementing an actual adaptive system poses many additional challenges, such a system has not been implemented as part of this thesis. Instead we again execute a post-analysis, but this time on the data gathered during the second experiment.
Once all the data of the second experiment was gathered, it was replayed and a simple scoring system was applied to judge the decisions an adaptive recommender *could* have taken given the data gathered during the experiment.
Using this approach most of the complicating factors concerning adaptive systems are taken out of play, while it is still possible to distil the potential effectiveness of such a system. An overview of the post-analysis procedure is provided in *Appendix H*.

The four contextual factors, as defined in *Chapter 5.2*, are taken into account. These four contextual factors are quickly summarised below.

1. Geographical Origin
   The first context classifies visitors based upon their geographic origin. During the second experiment both the country and continent of origin were taken into account. For the analysis executed in this chapter only the continent of origin is used.
2. Google Query
   The second context is concerned with the search-query a visitor enters on Google. This context is based upon a simple analysis of the query and classifies the visitor based on the outcome of this analysis. Details on this classification procedure are provided in *Appendix 1*.
3. Academic Discipline
   The third context considers the academic discipline of the Master's programme viewed by the visitor. The MastersPortal database contains a hierarchical list of academic disciplines; each programme is assigned to at least one discipline. Nine disciplines are at the highest level of the hierarchy. These are taken into account for the subsequent analysis.
4. Screen Resolution
   The final context looks at the screen resolution of the visitor. This tells us less about the interests of the visitor, but it is an important factor in determining the visual experience the MastersPortal.eu website offers to the visitor.

An additional fifth context, the type of browser used by the visitor, was taken into account initially. This context provided only a few interesting observation and similar observations were also apparent within the other contexts. As such this fifth context was dropped from the discussions in this chapter.

The goal of this chapter is to determine whether implementation an adaptive system is possible in the future. The intention is to show that enough data can be gathered and stability arises quickly enough for an adaptive system to steer recommendations in a useful manner. A further goal is to provide evidence on the fact that introducing *context-awareness* has a meaningful influence on the system as compared to only taking into account an overall perspective.

7.2 Scoring System

The adaptive recommender system envisioned as the final result of the analysis in this chapter starts out by weighting the results of the content-based and collaborative recommender approaches equally. Each visitor is classified based on the four contextual factors described previously. If the adaptive system detects either of the recommender approaches to perform significantly better than the other, this approach is given additional weight.

To ascertain the feasibility of an adaptive recommender system, a simple scoring mechanism is applied during this thesis. In this simple implementation, a single preference score is computed to indicate the level of preference towards either of the two recommender approaches. This score should provide insight into the behaviour of the proposed system when exposed to the experiment's dataset and as such, real-world visitors of the MastersPortal.eu website.

In an actual adaptive system, the preference score could be used to assign more display slots to the preferred recommender. Further iterations of the adaptive recommender system should implement more advanced weighting schemes though. By not simply computing a single preference score, but by combining the top results from both recommender approaches in a weighted manner, better adaptive recommendations are possible.

A schematic overview of the scoring process applied during this analysis is provided in *Figure 28*. Each time a session is started the adaptive system attempts to modify the recommendation presented based upon the contextual categories of the visitor at hand. To do so it classifies the visitor based upon its contextual factors and retrieves the past preference for each of the contextual categories thus created.

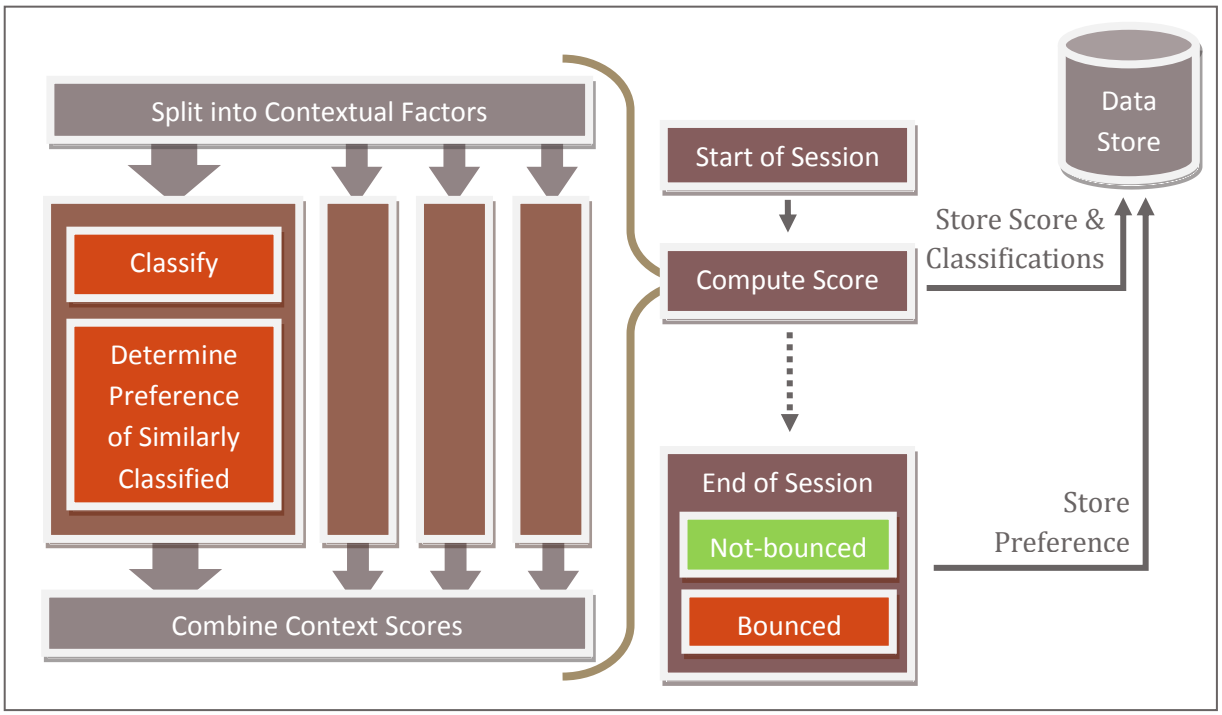


Figure 28: Schematic overview of the scoring process

Each of the four contextual factors has an equal weight in determining the overall score. A context is either in favour of the content-based recommender, in favour of the collaborative recommender or undecided. To arrive at a final preference score, the average of the scores of each of the individual contextual categories of the visitor is computed. This average thus score ranges from 100%, fully favouring the content-based recommender, to -100%, fully favouring the collaborative recommender.

The score for each separate contextual factor is based upon the same hypothesis test as used in *Chapter 6*. The bounce-rate for both the content-based and collaborative recommenders is approximated normally, after which Welch's t-test procedure is executed.
If the certainty one of the two recommender approaches has a lower bounce-rate than the other passes 99%, the context is counted as voting in favour of that recommender approach. Otherwise, the context's vote is undecided. A formalised representation of the scoring procedure is provided in *Figure 29* below.

Overall score: $S = \sum_{v \in V} S(v)$

Visit score: $S(v) = \sum_{c \in C} S(v, c)$

Context score:

$$\forall c \in C, v \in V: S(v,c) = \begin{cases} 1 \; if \; p\left(BR_{v,c,collaborative} > BR_{v,c,content}\right) \geq 0{,}99 \\ -1 \; if \; p\left(BR_{v,c,context} > BR_{v,c,collaborative}\right) \geq 0{,}99 \\ 0 \; else \end{cases}$$

Bounce-rate: $BR_{v,c,r} = Bounce\text{-}Rate\ of\ all\ visits\ prior\ to\ v;\ classified\ in\ context\ c;\ exposed\ to\ recommender\ r;$

With: $V = \{All\ Visits\ to\ MastersPortal.eu\}$

$C = \{Contextual\ Factors\}$

Figure 29: Formalised representation of the scoring procedure

As with all previous analyses, only participant arriving from Google are taken into account. Each time a visitor is referred by Google, the adaptive system makes a decision. All previous recommendations recorded for visitors referred by Google are used as a basis for the decisions. If a visitor starts a new session, by not visiting for at least sixty minutes, a new decision is made.
There is thus an implicit sixth contextual factor in play: Visitors need to be referred by Google. It is important to realise that in an actual adaptive system this piece of contextual information will need to be taken into account as a separate factor. For now, only visitors coming in from Google are taken into account; all other visitors are discarded right away by the scoring system.

7.2.1 Contextual Factors with Multiple Values

Two of the four contexts considered in this chapter, the “Google query” and “academic discipline” context, can classify visitors into multiple groups simultaneously. They are not mutually exclusive; something which the other two contexts are.

To take the example of the “academic discipline” context: A Master’s programme can be linked to multiple academic disciplines. If a visitor is classified using this Master’s programme, the visitor would be classified into multiple groups.
In these situations the scoring system does not compute a score for the visitor. This would introduce additional complexities that I want to avoid during this exploratory study.
Although visitors classified into multiple groups receive no score, their sessions are taken into account when calculating the scores for subsequent visitors. Especially for the “academic discipline” context this leads to 30% additional data being available for the scoring process.

In the graphs for both of these contexts, a fourth line is present. The “Sessions (Voting)” line indicates the number of sessions that provide information used in by scoring process. The number of actual scores computed is indicated by the “Sessions” line.

7.3 Overall Score

In this section the overall result of applying the scoring system on the data gathered during the second experiment is discussed. In addition this section provides an introduction to the discussion of the various contexts presented in the remainder of this chapter.

The graph in *Figure 30* shows the result of applying the scoring system without taking the contextual factors into account. All visitors are considered to be part of the same contextual category. The score for this categorisation is computed by simply taking into account all previous visitors. The result of this analysis establishes a baseline to which we can compare the results of an adaptive system when taking into account the various contexts.

The graph below consists of three lines which together provide for an optimal interpretation of the scoring result. The first line, "Score", is the actual result of the scoring system. It is the decision made by the scoring system for the current visitor based upon all similarly classified previous visitors. Its implementation is discussed in the previous chapter.

Secondly, "Baseline" represents the expected preference of the visitor if he or she would prefer the recommendation presented. This line shows the relative portion of visitors presented with each of the recommenders. The average of this line over the entire two week interval is close to zero; it is effectively an "over time" graphing of the distribution of experiment participant amongst the two experiment groups from *Chapter 6*.

Thirdly, the "Sessions" line represents the number of sessions available to the scoring system. This provides an indication of how the amount of data available influences the stability of the system.

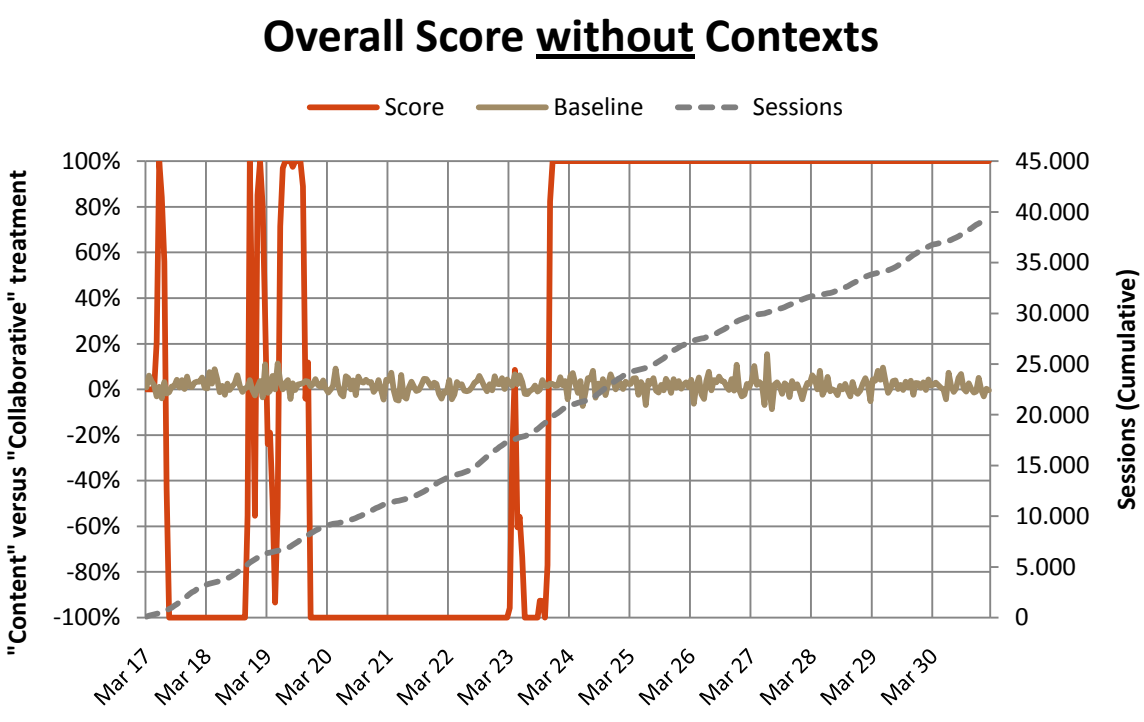


Figure 30: Overall score without taking the four contexts into account

Looking at the score in *Figure 30* we see an unstable situation for the first three days. After the third day, the adaptive system prefers the collaborative recommender for three days. Subsequently, preference switches towards the content-based recommender and remains there for the remainder of the experiment.

Considering the overall score from *Figure 30* we see it reaches the same conclusion as those discussed in Chapter *6*. There is no statistically significant distinction to be made between the performance of the content-based and collaborative recommenders. Although, it does appear as if the content-based recommender performs slightly better.
Just like in *Chapter 3,* we also see the relative performance of the two recommenders is not stable over time. Without taking the contexts into account, an adaptive system can thus still add value: The overall preference changes over time.
Considering the baseline in *Figure 30* we see a fair and stable distribution amongst the two experiment groups. This is again in line with the results of the second experiment. The number of sessions increases linearly, the slight variation is caused by the daily visitor trend.

The overall result of the scoring system when the four contexts are enabled is graphed below in *Figure 31*. This figure shows the scores as computed by the adaptive system according to the full contextual scoring procedure as described in *Chapter 7.2*.

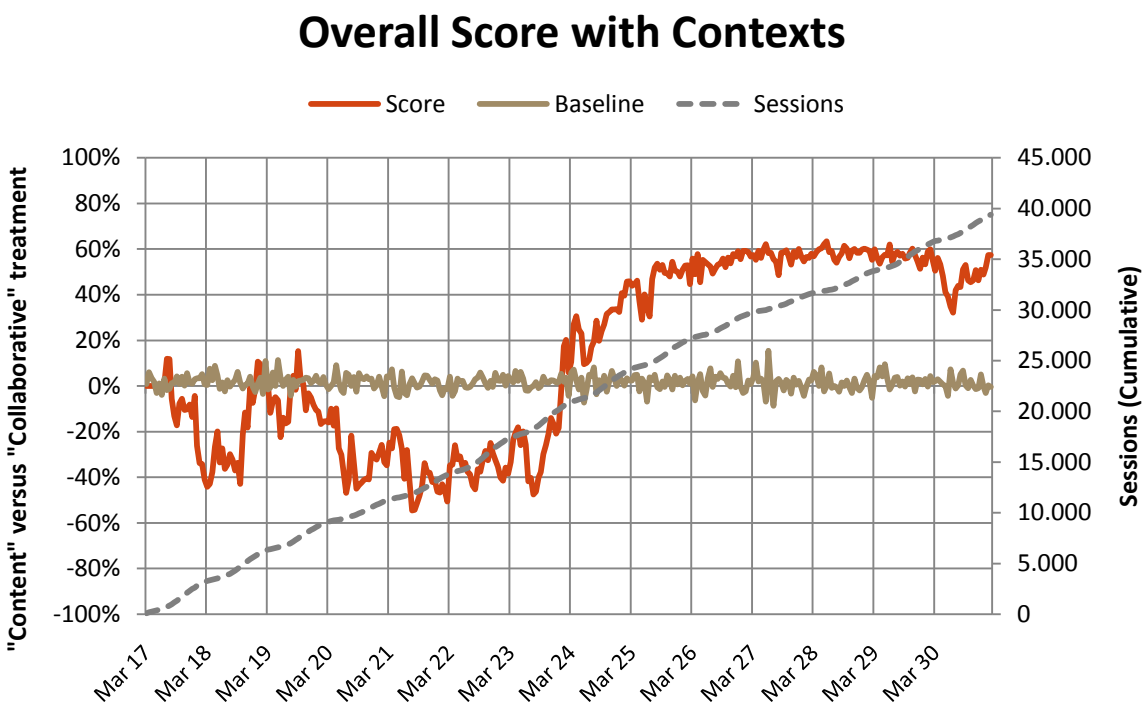


Figure 31: Overall score when taking the four contexts into account

The overall score is less outspoken this time around. Even though the graph follows the pattern established in *Figure 30*, it never reaches the full preference for either of the recommenders. These results show that when taking the contexts into account we find groups of participants who do not agree with the decision made by purely looking at the overall score from *Figure 30*. Introducing contexts clearly influences the decision made by the adaptive system.
We also see that the score stabilises as time progresses. The more data becomes available, the less outspoken the spikes in the score become. At the end of the two week period there is a relatively stable 60% score towards the content-based recommender.

To better understand the effects of the contexts, the next section looks at the scores for each of the contexts independently.

7.4 Context Scores

To gain more insight into the effects of the different contexts, the analyses presented in the previous section are again executed, but now separately for each context. An overall score for each context is computed and all the visitor categories within the context are analysed separately. This provides a good indication of how each context influences the overall score.

For each of the four contexts discussed in this chapter, a graph is provided that displays the score of the context. This graph allows the effect of the single context on the overall score to be interpreted. The graphs are similar in setup to those presented in the previous section.
Furthermore, a table with an overview of the result for each category within the context is provided. The results listed in this table are based upon an interpretation of the graphs provided for each contextual category in *Appendix J*.
In the tables the final number of sessions for each category is provided. In the case of the "Google query" and "academic discipline" contexts, the number of scores computed is lower than the number of sessions used for data gathering. For these contexts, two numbers are provided. In the graphs for both of these contexts, a "Sessions (Voting)" line is also included.
Finally, each table contains a "Stable" column which is used to indicate if and when a contextual category reached stability. The decision on stability is based upon an interpretation of the graphs referred in the table. It should be taken as an indication of stability, not as hard statistical evidence.

7.4.1 Geographical Origin

The first context discussed is the geographic origin of the visitors. Instead of looking at both continents and countries, as was done during the second experiment, the current analysis only takes the continent of origin into account; country information is discarded.
Only a small number of countries provide enough participants to be considered significant. Taking into account the country aspect of the context would thus add a lot of complexity, without adding much in the form of additional results.

The overall score for the geographic origin context is displayed in *Figure 32* on the next page. The score is instable during the first week of the replay. Afterwards the context stabilises and shows a strong preference towards the content-based recommendation.

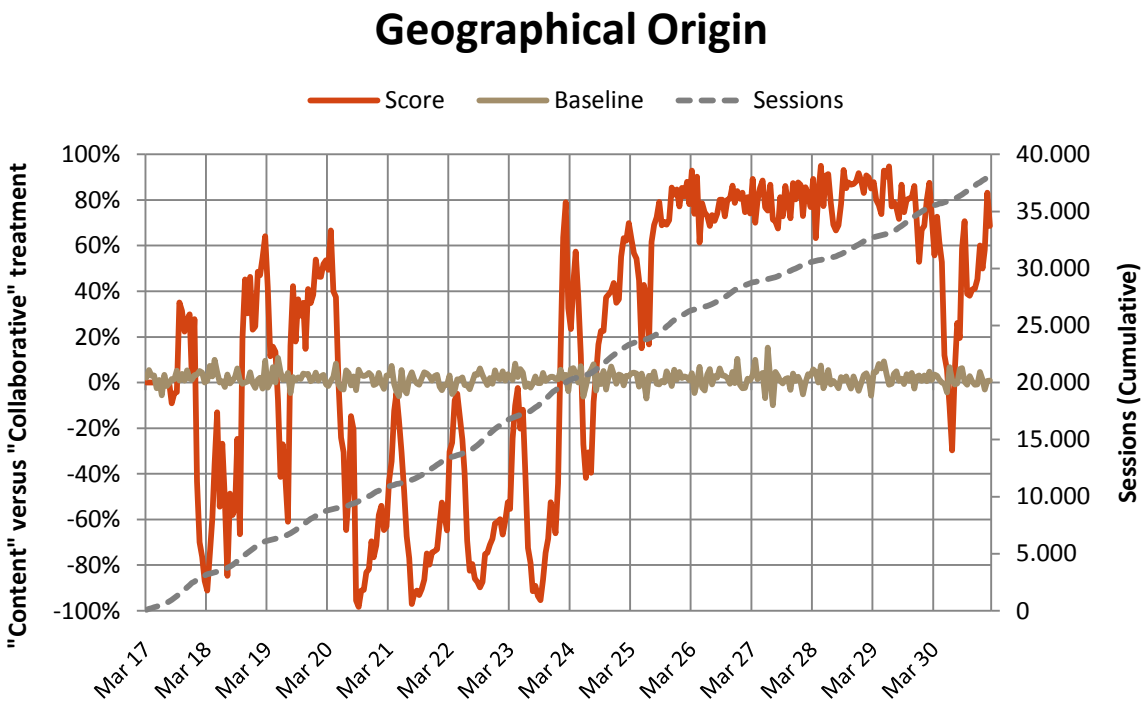


Figure 32: Overall score for the "Geographic Origin" context

For 1.321 participants, the scoring system was unable to detect a geographical origin. The sessions started by these participants were discarded. Due to the fact that this only involved a fraction of the total number of participants, we this exclusion has no effect on the stability of the baseline provided in the above graph.

In the unstable period at the start of the graph, a clear daily pattern emerges. Once the score stabilises, this pattern disappears. The pattern is especially clear between the 21st and 23rd of March up to the extent that it is visible in the overall score graphed in the previous section. Furthermore, the dip we see on the 30th of March in the overall score appears to be caused by the geographic context.

In *Table 20* an overview of the results for each of the categories within the geographic context is provided. In *Appendix J1* a detailed graph is provided for each of the continents listed in the table.

Continent	Sessions	Preference	Stable	Appendix J1
Africa	2.819	Collaborative	2 Days	Figure 60
Asia	9.708	Undecided	-	Figure 61
Europe	18.226	Content	7 Days	Figure 62
North America	5.822	Content	2 Days	Figure 63
Oceania	804	Undecided	-	Figure 64
South America	697	Undecided	-	Figure 64

Table 20: Overview of results for the "Geographic Origin" context

Both Africa and North America quickly stabilise towards the collaborative and content-based recommenders respectively. This is in accordance with the results presented in *Chapter 6*.

Europe reaches stability after seven days. Considering the large amount of sessions classified into Europe it is quite surprising that it takes this long. The overall stability of the geographical origin

context seems to be caused by Europe gaining a significant preference.
The second largest category, Asia, is hovering around indecision. Sometimes it prefers the content-based recommender, sometimes is prefers the collaborative recommender. Overall this leads to Asia having no clear preference, as is also indicated in *Chapter 6.4*. Looking at its performance over time we do see stretches of significant preference towards either one of the recommenders though.

Both Europe and Asia appear to be too large to act as effective categories. This is most likely caused by large number of nations and different cultures within these continents. There is not enough data to draw significant country-level conclusions, but as a result of country-level differences, the overall continent categories never stabilise.
This conclusion is reinforced by the fact that both Africa and the North America do reach stability quickly and with less data available. From experience within MastersPortal we know that education seekers from Africa and North America indeed have a more homogenous information need.

Oceania and South America need several days to meet the statistical assumptions required for Welch's t-test to be applied to their data. These two continents are also by far the smallest. Eventually, South America seems to favour the content-based recommender, whereas Oceania seems to favour the collaborative recommender. Looking at the dataset as a whole I am though reluctant to conclude either of these two continents reaches stability within the two week interval.

7.4.2 Google Query

The second context investigated in more detail is the Google query context. The overall score of the context is graphed in *Figure 33* below. Compared with the other contexts, the Google query context contains far fewer sessions. In the graph we clearly see the effect this has on the stability of both the score and baseline.

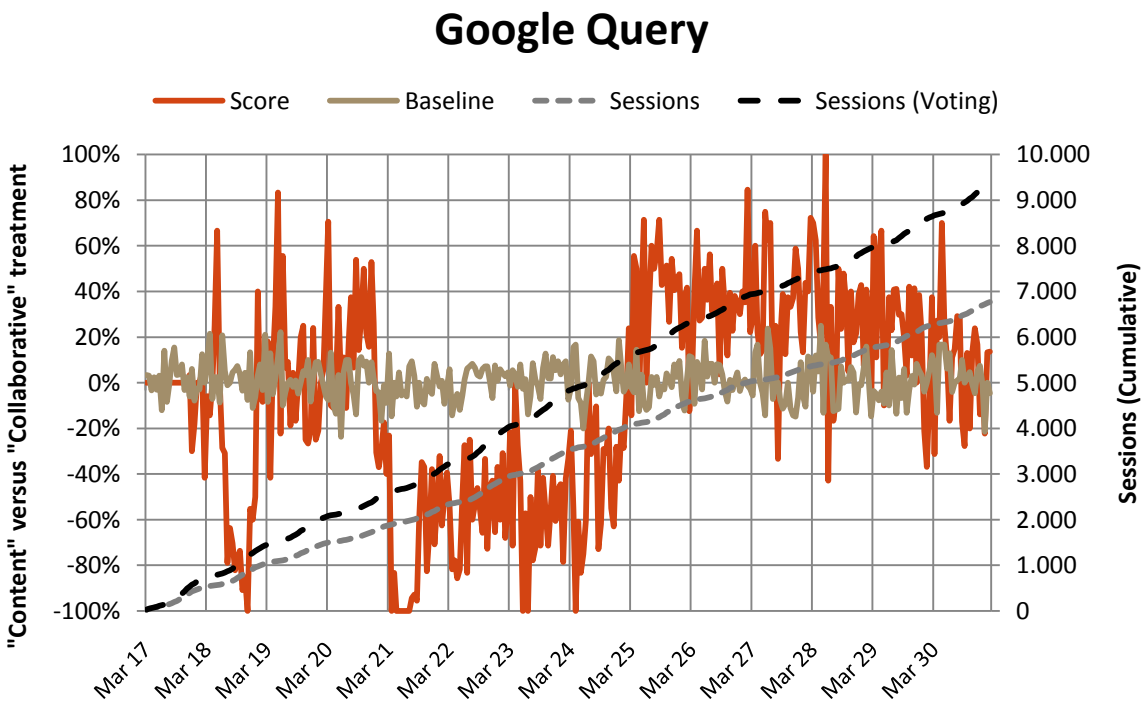


Figure 33: Overall score for the "Google Query" context

The score of the Google query context is less stable than that of the other contexts. This is caused by the relatively small amount of data available for this context. We do see a clear trend throughout the graph, but the actual score varies greatly. A second observation is that the score of the Google query context is more centred towards the zero line of the graph than the other three contexts.
This makes it an interesting context, as it divides participants more equally than the other contexts. By utilising the Google query, the maximum amount of visitors who would otherwise be presented with a worse performing recommendation can be provided with a better one.

In *Table 21* an overview of the results for each of the categories within the Google query context is provided. In *Appendix J2* a detailed graph is provided for each of the four query categories available.

Google Query		Sessions	Preference	Stable	Appendix J2
Interest in Master's	14.211	-	-	-	-
Europe	533	(871)	Undecided	-	Figure 66
Country	1.879	(2.629)	Collaborative	2 Days	Figure 67
University	240	(1.292)	Undecided	-	Figure 68
Title Match	2.838	(4.670)	Content	8 Days	Figure 69
No Interest	25.186	-	-	-	-

Table 21: Overview of results for the "Google Query" context

All categories for the Google query context are a combination of the "Interest in a Master's Degree" classification with one of its four sub classifications.
As can be seen from *Table 21* around one third of all sessions show a general interest in a Master's degree. Within this classification, again around one third of sessions are classified into a single of the four sub classifications. For these approximately 5.000 sessions a score was computed during the analysis of the Google query context. In braces behind the session counts is the number of sessions used for data gathering. These are also listed as "Session (Voting)" in *Figure 33*. This number includes sessions which were classified into multiple groups.

The two most important categories in this context are "interest in a country" and a "title match". The country interest category reaches stability after about two days. The title match classification reaches stability after eight days.
The preference towards the collaborative recommender for the country interest category is also found in the results presented in Chapter *5.2.2*. The title match preference for the content-based recommender is visible in *Chapter 5.2.2*, but nowhere near significant. The most likely cause of this discrepancy is the fact that more relevant data was gathered for the Google query contextual factor during the second experiment.

Both the "interest in Europe" and "interest in a university" classifications contain a limited number of recommendations. Because of this, they both need three days to meet the statistical assumptions required for Welch's t-test to be applied to their data. Afterwards both remain unstable for the remainder of the experiment. Both categories are simply not large enough to provide any added value to the Google query context.

7.4.3 Academic Discipline

The third context investigated in more detail is the context concerning itself with the academic discipline of the Master's programme viewed by the visitor. The overall score of the academic discipline context is graphed in *Figure 34* below. The context is rather unstable during the first week,

but eventually reaches a stable state. The results of this context provide a clear picture of how both the score stabilise as more data becomes available.

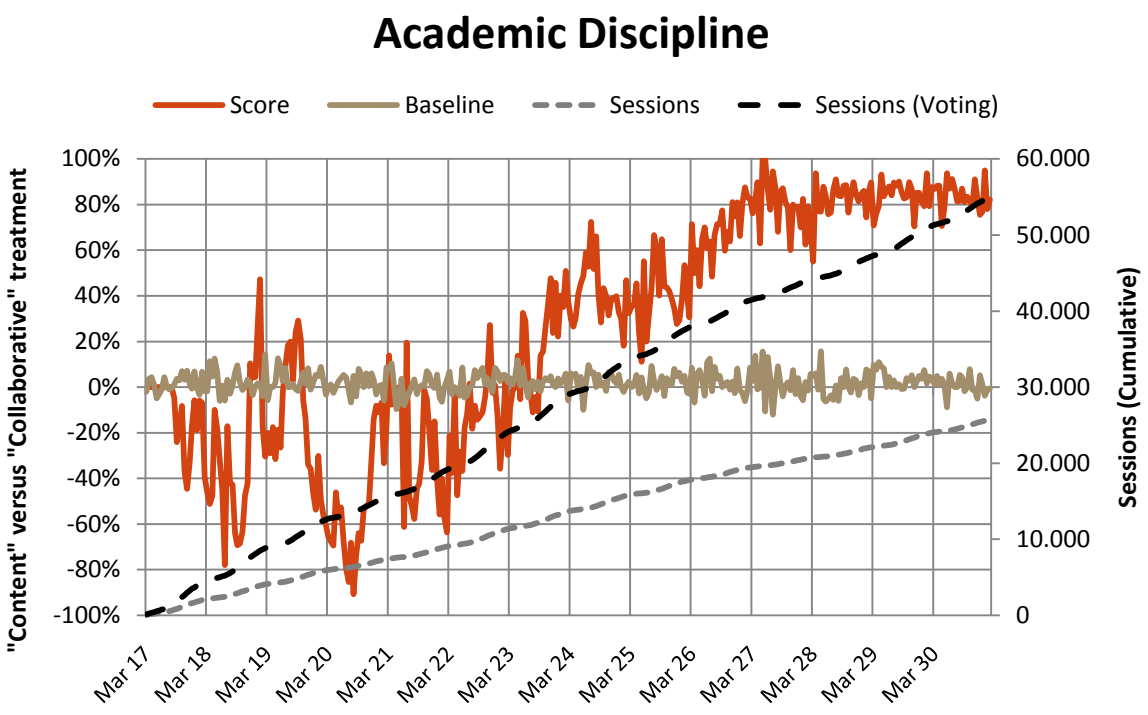


Figure 34: Overall score for the "Academic Discipline" context

Looking at the graph in *Figure 34* we see that after stabilising, most of the discipline categories favour the content-based recommender.
Just as with the Google query context, the academic discipline context also classifies some sessions into multiple groups. These sessions were again not scored, but they were used for data gathering. This is again displayed in the graph through the inclusion of a fourth line, "Sessions (Voting)" which indicates the number of sessions used to gather data.

In *Table 22* an overview of the results for each of the categories within the academic discipline context is provided. In *Appendix J3* a detailed graph is provided for each of the nine top-level discipline categories available.

Academic Discipline		Sessions	Preference	Stable	Appendix J3
Law	1.142	(2.000)	Collaborative	3 Days	Figure 70
Engineering & Technology	6.133	(11.054)	Content	6 Days	Figure 71
Humanities & Art	2.727	(4.623)	Content	7 Days	Figure 72
Life Sciences, Medicine & ...	3.162	(5.172)	Content	6 Days	Figure 73
Natural Sciences	872	(4.477)	Collaborative	1 Day	Figure 74
Applied Sciences, ...	2.118	(5.385)	Undecided	-	Figure 75
Social Sciences	3.476	(8.442)	Undecided	-	Figure 76
Business & Economics	5.952	(10.419)	Content	7 Days	Figure 77
Environmental Sciences	182	(3.522)	Undecided	-	Figure 78

Table 22: Overview of result for the "Academic Discipline" context

Most of the disciplines seem to strongly prefer a single recommender approach. They do not exhibit the switching behaviour that is present in the other contexts. As a result, the overall score for the academic discipline context shows less preference towards the collaborative recommender during the initial week of the replay.

The “Law” and “Natural Sciences” categories favour the collaborative recommender. Both of these disciplines are small when compared to the other disciplines in *Table 22*, but both quickly and strongly prefer the collaborative recommender. This same effect is visible in *Chapter 5.3.4*. During the second experiment, more data was gathered than available for the analyses presented in *Chapter 5*, leading to a significant result for both categories.
The fact that two categories strongly prefer the collaborative recommender, in opposition to the overall preference, indicates the academic discipline is relevant to the adaptive system.

Four categories favour the content-based recommender. All four of these categories are unstable during the first week of the replay. They start favouring the content-based recommender after about seven days. Apart from the “Business & Economics” category, a similar effect is observed for these disciplines as part of the analyses in *Chapter 5.3.4*.
Taking an overall perspective on the “Social Sciences” category indicates it is undecided. A more detailed look at the category shows it has a significant preference quickly, but due to the preference switching halfway through the replay the final result is undecided.
The “Applied Sciences” discipline is unstable throughout the replay. It appears as if the discipline stabilises in the last few days of the replay. “Applied Sciences” is, from a conceptual perspective, the broadest discipline in the MastersPortal database. This might well explain its instability throughout the replay. The discipline contains many applied courses spanning most of the other disciplines in the MastersPortal database. Due to its broad conceptual content and relatively small number of programmes stability is not reached quickly.

The “Environmental Sciences” category is an exceptional case. Initially the MastersPortal database contained eight top-level disciplines. “Environmental Sciences” was added later on. As a result, the discipline has a small number of programmes assigned to it exclusively. Many Master’s programmes were added to the discipline at a later point; they were not removed from their original disciplines. Although it appears the category might favour the content-based recommender, it is interpreted as undecided due to the reason noted above.

Concerning significance we see an interesting effect: Some of the larger categories take longer to stabilise than the smaller categories. This strongly points to the fact that some of the discipline categories are not focussed enough.
Generally speaking, the academic discipline context does add value to an adaptive system utilising contextual factors. Several of its visitor categories behave differently from the overall result and as such can be used to optimise the recommendation presented.

7.4.4 Screen Resolution

The fourth and final context investigated in more detail is the screen resolution context. The overall score for the screen resolution context is displayed in *Figure 35* below.

Looking at the overall impact of the screen resolution context, especially after it stabilises, we see it does not add much distinctive power to the adaptive system. Participants are so equally distributed amongst the four categories that their behaviour is the same as that of the system without taking the contexts into account, graphed in *Figure 30*.

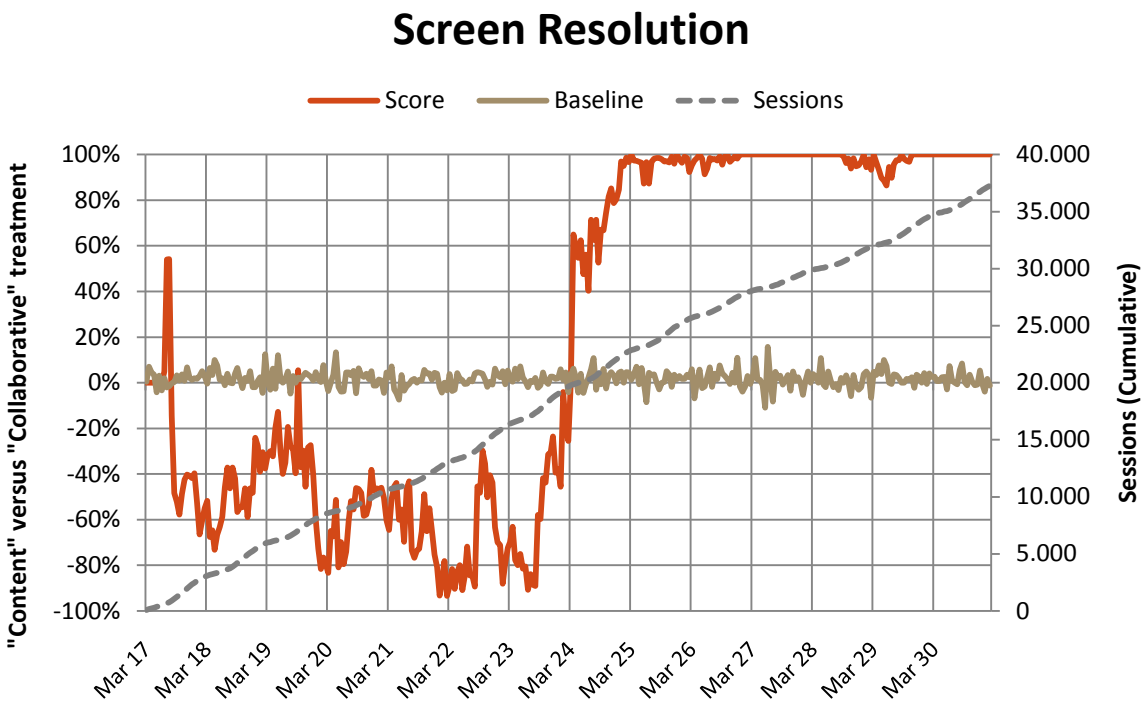


Figure 35: Score and baseline for the "Screen Resolution" context

The results of the screen resolution context provide us with an interesting observation: They point to the fact that the amount of "programme recommender" box (as shown in *Appendix C*) visible without scrolling does not have an impact on visitor behaviour. Interested visitors apparently fully explore the page before navigating elsewhere. As a result, it might prove useful in the future to consider factors apart from the screen resolution when determining the number of programmes to recommend, or how to position the recommendation.

In *Table 23* an overview of the results for each of the visitor categories within the screen resolution context is provided. In *Appendix J4* a detailed graph is provided for each of the four screen resolution categories distinguished.

Looking at the four visitor categories in *Table 23* we see all of them stabilise and favour the content-based recommender. This observation is in line with the results presented in *Figure 35*.

Resolution	Sessions	Preference	Stable	Appendix J4
Tiny	2.322	Content	4 Days	Figure 79
Small	6.210	Content	8 Days	Figure 80
Medium	21.404	Content	7 Days	Figure 81
Large	6.545	Content	7 Days	Figure 82

Table 23: Overview of results for the "Screen Resolution" context

When we compare the results of the “Medium” category with the overall results in *Figure 30* we see a striking resemblance. After some initial instability the “Medium” category prefers the collaborative recommender; after seven days preference switches towards the content-based recommender. The distribution of participants within this category is so alike the overall participant distribution that the “Medium” category provides nearly the same result. It thus provides little added value to the adaptive system.
The “Small” and “Large” categories both behave similarly; both categories contain an equal numbers of participants. During the initial week of the replay the categories are unstable. After this both prefer the content-based recommender. As both categories are much smaller than the “Medium” category, the initial instability is most likely caused by a lack of data and not by an inherent instability. As such, I would consider the “Small” and “Large” categories too as providing little added value to adaptive system.

Finally, “Tiny” is the only category which does not closely confirm to the overall result. It stabilises towards the content-based recommender several days ahead of the other categories and also well ahead of the overall result. Due to its relatively small size, the effect is subtle in comparison with the overall results of the screen resolution context.
This observation serves to somewhat nuance my initial conclusion: It appears that screen resolution is not an important factor in recommender performance, but visitors at the low end of the resolution spectrum do seem to be influenced by it. When *and* if this effect is a factor to take into account is something which cannot be determined from the current analysis.

My general observation with regard to the screen resolution remains that it has little added value. Classifying participants based on their screen resolution does not provide any additional information that can be used to provide a more suitable recommendation.

7.5 Conclusions

Looking at the overall results of the adaptive recommender system when no contexts are used, we see it initially favours the collaborative recommender. After several days preference switches towards the content-based recommender. When contextual visitor categories are taken into account a similar, but less outspoken, pattern is visible.
When using contextual information, the adaptive system closely follows the overall trend, but it is able to add nuances based upon the contextual categories of each visitor. Together with the effects identified within each context independently, the overall conclusion of this chapter is that adding contextual information to our adaptive system has the potential to improve its performance.

Judging by the overall result of the adaptive system, it would seem that with the current number of visitors to the MastersPortal.eu website, a week of visitor data is required for the system to reach stability. Although the overall results follow the non-contextual pattern closely from the start, we see a lot of variability in the first week. As time progresses, the variation becomes less outspoken. This conclusion is affirmed by looking within the contexts. Many of the contextual categories require at least a week to reach a stable state.
Looking at the different contexts in more detail reveals that large contextual categories are not by definition more stable than the smaller categories. As such, the number of sessions in a visitor category is not the most important factor influencing its stability. Focus within the context is also of high importance. This is especially clear within the geographical context and the academic discipline context. Here, many of the larger categories require more time to stabilise than the smaller ones. This is caused by them being relatively broad in their classification of visitors.

The geographical context allows for several of its categories to reach a stable preference towards one of the recommender approaches. As such it provides an added value for the adaptive recommender system. As noted above, some of the geographical categories are too broad. Especially Europe and Asia seem to be unfocussed. Due to the large number of countries in these continents it would be beneficial to introduce sub-regions. Classifying up to country-level is also possible, but this would result in many of the smaller countries not stabilising.
The Google query context is also relevant as it divides participants most equally between the two recommender approaches. It is at the same time a difficult context though, as a good analysis of the query string is required. The current quick analysis does not suffice in the long run as it discards too much relevant information. As such, the Google query context contains the smallest visitor categories, half of which do not contain enough data to stabilise at all.
The academic discipline context is a relevant and stable context. It contains only a few undecided visitor categories. Just as with the geographical origin context, the academic discipline context does contain some unfocussed categories. The MastersPortal database contains a second level of disciplines, which could be used to refine some of the current classifications made. Here also goes that some of the sub-disciplines are simply too small to be of use.
Finally, the screen resolution context behaves very similar to the overall result when no contexts are considered. This leads to the conclusion that this context is not particularly interesting for the adaptive system. It does not influence the overall behaviour of the system at all. As such I would consider it best to not further take this context into account.

8. Conclusions and Implications

Based on the analyses presented in this thesis, several conclusions can be drawn. This chapter presents the main conclusions of the thesis and their implications. It also provides an inroad into the final chapter of this thesis concerning future directions for MastersPortal and the system developed and discussed throughout this thesis.

8.1 Recommender Experiment

The main conclusion of the recommender experiment is that adding a programme recommendation to the MastersPortal.eu website has an undoubtedly positive effect on the bounce-rate of visitors coming from Google. A drop in bounce-rate of over 90% is achieved when we compare the best performing recommender approaches against the situation without a recommendation. Furthermore, the perceived relevance of the recommendation provided also appears to be an important factor.

Comparing the irrelevant (c.q. random) recommendation against the situation with no recommendation, we see no significant difference in bounce-rate. This indicates that simply adding some links under the guise of a "Related Programmes" feature has no effect. In order for the recommendation to have a positive result, it needs to provide visitors with actually relevant recommendations, based upon the Master's programme presented to them.

The baseline recommendation, which uses academic discipline hierarchy to determine relevance, offers a significant drop in bounce-rate when compared to the situation without a recommendation. When we compare this "simple" recommendation against the two automated approaches we see that both of these have double the effectiveness of the baseline recommendation.

The two automated recommenders, one content-based and the other collaborative, perform equally well under the current statistical assumptions. If we slightly relax the assumptions, we see that the content-based recommendation outperforms collaborative recommender by a small margin.
As the statistical assumptions used are quite strict, results do seem to indicate the content-based recommendation outperforms the collaborative recommender. The collaborative recommendation should not be discarded though.
The results of the contextual analysis executed after the first experiment make it clear that contextual factors have an influence on the relative performance of the content-based and collaborative recommenders. Several contextual categories show a significant preference for the collaborative recommendation approach.

As such, further development of the collaborative recommender seems to be prudent. The collaborative recommender provides a relevant recommendation which can be combined with the content-based recommendation to further enhance performance.

Finally, it is certainly beneficial to invest time and effort in implementing the automated and technologically more advanced recommender approaches. Both the content-based recommendation and the collaborative recommendation provide a significant additional reduction to the bounce-rate when compared to the baseline recommendation.

The implementation of the baseline recommendation, based on manual categorisations into an academic hierarchy, requires much less effort. Both automated recommenders do provide such an added performance benefit that this offsets their higher implementation costs.

An added benefit in case of the content-based recommender is the fact that much of its technology can be easily repurposed to create a better keyword search-engine than the one currently available on the MastersPortal.eu website.
As the MastersPortal expects to invest heavily in community-based features in the future, it will also become easier to gather the user-behaviour information required for the collaborative recommendation to perform better. The additional costs of implementing this recommender will for a large part be covered by the general development effort of the future MastersPortal.eu website.

Implications

Based on the results of the first experiment, implementing either the content-based or collaborative recommender reduces the bounce-rate for visitors referred by Google from around 90% to nearly 82%. This drop in bounce-rate almost doubles the number of visitors referred by Google who stay on the MastersPortal.eu website after their initial page view.
When we take the visitor statistics from March 2010 for the MastersPortal.eu website, implementing either recommender will *convert* 18.000 additional visitors. This adds 72.000 programme views to the monthly gross total, increasing it by almost 14%. As MastersPortal's current revenue model is largely based around the number page impressions generated, this increase in programme pages viewed leads to sizeable growth in revenue.

For the time being, the content-based recommender is implemented on the MastersPortal.eu website. Although initially it was technically somewhat more challenging than collaborative recommender, most of the work has been done upfront. Now that the recommender is fully implemented, it can be maintained with relatively little effort.
This as opposed to the collaborative recommender: It needs to be fed a continuous stream of rigorously filtered behavioural information to provide good recommendations. Nonetheless, the future will see further enhancements in the direction of collaborative recommendation. Certainly when considering that MastersPortal aims to further develop its community features in the near future. Some of the additional potential of the collaborative recommender is discussed in the next, and final, chapter of this thesis.

8.2 Contextual Factors

From the experiment and the analysis concerning the contextual factors stems the conclusion that simply looking at the overall performance for each recommender type does not do complete justice to its specific merits.
Overall we see that the content-based recommender slightly outperforms the collaborative recommender. When we take different contextual categories into account, much greater and some opposite performance differences are observed.
This provides an indication to the fact that any optimal recommender system will need to combine multiple, technologically different, approaches to provide visitors with the best recommendation possible.

During the second experiment, the existence of contextual influences was confirmed. In more than a couple of situations, contextual categories were found in which the collaborative recommender significantly outperformed the content-based recommender.
During the experiment, specific attention was paid to the visitor's geographical context. Visitors were classified both at the continent and at the country level. From the experiment we conclude that currently visitors from Africa and Germany strongly prefer a collaborative recommendation, in opposition to the overall result of both experiments.

The second experiment also unearthed contextual influences that were not present during the post-analysis executed on the data of the first experiment. This provides a second important conclusion: The influence of the contextual factors is not necessarily fixed over time. Between the first experiment and the second experiment, two months passed and significant changes in the influence of the contextual factors occurred.
As such, classifying visitors based on the contextual factors in order to optimise the performance of the recommender system will require a dynamic approach. Although this approach does not have to be automated, the rules according to which visitors are classified will need to be reviewed at certain intervals.
A further conclusion is that the contextual categories need to be carefully selected. Both the post-analysis and the second experiment show that some categories, especially within the geographic and academic discipline contexts, are too broad. They do not differentiate at all from the overall results. Choosing the right contextual factors and their categories requires a good deal of research and interpretation in advance of any implementation.

Implications

The contextual categories of a visitor matter greatly if we want to further optimise the overall performance of the recommender system. It is difficult to predict the influence of each context in advance and the influence of the contexts changes over time.
Any system utilising the contextual information uncovered during this thesis will need to be, up to some extent, dynamic. It needs to be tuneable to cope with the changing environment operates in.

As a result of this an additional analysis, not planned at the start of this thesis, was executed. In an attempt to lay the groundwork for an adaptive recommender system the effects of several different contextual factors were analysed over the course of several weeks. The conclusions stemming from this analysis can be found in the following section of this conclusion.

8.3 Feasibility of an Adaptive Recommender System

At the core of the analysis concerning the adaptive recommender system is the goal to determine whether it is possible for an automated system to pick up on differences in preference made apparent by classifying visitors along the lines of several contextual factors. After identifying these preferences, a future adaptive recommender system can modify its recommendations based upon them. As the analysis was executed offline, no modifications to the actual recommendation presented to the visitor were made. The goal of the system designed for this thesis was simply to show the data available provides enough statistically relevant information to sustain an adaptive recommender system.

In light of this goal, a simple scoring system was implemented that combined the preferences of four contextual factors into a single overall score. The potential relevance of the adaptive system is indicated by the fact that its overall score reaches a stable state after about a week is made available to it.
After eight days of instability, the overall score of the adaptive system stabilises with a 60% preference towards the content-based recommender. This is similar to what we expected, judging by the result of both experiments executed during this thesis.
The actual "60% preference" is not all that relevant at the moment. It is merely an indication of which fraction of the contexts favour the content-based recommender. In a final implementation there will most likely be individual weights for each context.

A further conclusion is that in an "adaptive" recommender which always presents the results of a single recommendation approach, certain groups of visitors will always be provided with a recommendation that does not perform optimally for them. As such, there is certainly potential for a real adaptive system.
Looking into the different contexts we see several categories that prefer the collaborative recommender over the content-based recommender. This indicates a future adaptive system will be able to optimise the overall performance by tweaking the recommendations it presents based upon the contextual categories of the visitors.

Finally, the results of the contextual factors analyses lead to the conclusion that statistical relevance alone is not enough to base changes in an adaptive system on. The stability of the statistical relevance over time is also of great importance.
This is something MastersPortal will need to take this into account when further developing its adaptive system, or any system which modifies the recommendations based upon statistical evidence. The system cannot simply switch when statistical significance is achieved; it needs to wait for a both significant and stable state to set in. This conclusion might be obvious, but it seems like something which is easily overlooked nonetheless.

Implications

The analysis discussed in the previous section show that, given sufficient data, an adaptive recommender system is able to reach a stable state within a short period of time. My expectation is that this stable state, which mixes the content-based and collaborative recommenders, performs better then when the recommender system confines itself to either approach. Further research is though required to solidify this statement.

The contextual factors and their categories need to be chosen very carefully. From the results of *Chapter 7* it is clear that improperly chosen contextual categories provide no new insights, despite the availability of a large amount of data.
In the results of the analysis we see for example that some of the academic discipline categories and visitors from Europe and Asia have a large number of participants, but do not significantly deviate from the overall pattern without contextual information. As such further investigation into the best contextual factors for MastersPortal is required.
As a starting point I would recommend a system which uses the Google query, academic discipline and geographical origin contexts to base its adaptations on. The Google query context is somewhat unstable, but it does offer important information on the visitor's interests.
For the Google query context to be useful in practice, the technique used to classify its participants will need to be enhanced. This allows for more participants to be properly classified, which in turn leads to a more stable behaviour. If this is achieved I am confident that the Google query context adds a lot of value to the envisioned adaptive system.

As noted in the previous sections of this conclusion, MastersPortal will in the near future invest more effort into community-based features. With this, it will be possible to compile more a complete profile of its visitors.
The information gathered in these profiles needs to be combined with the contextual factors identified in this thesis to come to the best possible visitor categories. These categories can increase performance of both the collaborative recommender and any future adaptive recommender system. By combining the contextual factors with user profiles it will also be possible to better infer the preferences of new and unknown visitors based purely on their contextual categories.

Much additional work is required before the adaptive system as discussed in this chapter can be implemented on the MastersPortal.eu website.
The results of this thesis simply point to the fact that dynamically modifying recommendations based upon visitor characteristics has the potential to improve performance. In the next and final chapter of this thesis, the future prospects and potential issues of such an application are discussed in more detail.

8.4 Summary of Contributions

The technologies implemented as part of this thesis and the results of the experiments executed have made several contributions to the way the MastersPortal.eu website currently works and how it will evolve in the future. This chapter outlines these contributions and their implications to the MastersPortal's business in general.

8.4.1 Programme Recommender on MastersPortal.eu

The most visible contribution of this thesis is the inclusion of a recommender system for Master's programmes on the MastersPortal.eu website. Shortly after analysing the results of the first experiment, the content-based recommender approach implemented as part of this experiment was rewritten into a fully functional recommender system and deployed on the live MastersPortal.eu website.

This recommender system has been up-and-running since February of 2010. An example of the system in action is available on the page of the *Business Information Systems* Master's programme[10]. To find this programme search either for "Business Information Eindhoven" on the MastersPortal.eu website or follow the URL provided in the footnote.

MastersPortal has received positive feedback from both its university customers and student visitors in response to the introduction of the recommender system. Since the introduction of the system a clear drop in bounce-rate is visible for visitors coming from Google. This drop is in line with the results of the experiment.

The technology powering the recommender system can furthermore be easily repurposed for use in other parts of the MastersPortal.eu website; especially within its search-engine enhancements based upon the technologies developed are easily realised. In principle, the technologies developed during this thesis can help to improve both searching and navigational features in the same way as they helped to enhance MastersPortal's recommender system.

Improving the MastersPortal.eu search-engine is one of the first projects I will personally work on after completion of this thesis. The tf-idf vector space retrieval code written as part of the content-based recommender will be used to improve the MastersPortal.eu keyword search-engine.

8.4.2 Birth of an Experimentation Culture

Apart from providing evidence towards the positive effects of implementing a recommender system on the MastersPortal.eu website, the work done as part of this thesis has raised awareness on the practical value of executing online controlled experiments.

As noted in the background section, an important guideline at Amazon is that "data trumps intuition". During my literature research prior to this thesis I have come across many examples in which a solution favoured by domain experts is significantly outperformed by a solution these experts deemed inferior. Decision on how to optimise website layout and their features at companies such as Microsoft and Amazon are made based upon hard data, the results of online controlled experiments, and not upon (expert) intuition.

Concerning MastersPortal, questions often arise as to what would be the best way to implement a feature or whether implementing a feature is smart at all. The experiments executed as part of this thesis have shown that these questions can be answered through the use of online controlled

[10] http://www.mastersportal.eu/students/browse/programme/69/business-information-systems.html

experiments. The experimentation system deployed as part of this thesis provides the basis for a future integral experimenting approach within the MastersPortal.
This thesis provides information on how experiments are best constructed, executed and evaluated; both on a conceptual and on a practical level. With the experience gathered as part of this thesis, MastersPortal has a good indication of the kind of results it can expect from the application of online controlled experiments. We furthermore know how much and what kind of data is required and how stringent we should adhere to experimental assumptions.

8.4.3 Contextual Factors and Adaptation: The Future

The analysis of contextual factors and the additional effort that went into the creation of an adaptive recommender system opens up a clear future direction for the MastersPortal.eu website and the other initiatives currently under development by MastersPortal.

Firstly, the analysis of the different contextual factors has lead to a greater appreciation of the heterogeneous nature of visitors to the MastersPortal.eu website. It is of course obvious that a large geographical spread of visitors leads to a mixed population, but that the differences can be made clear so easily was not expected. As such the importance placed on recognising and utilising these differences within future developments has increased.
Within MastersPortal it has always been a goal to work towards the creation of a student community and a set of community-based features. The results of the contextual factors analysis only serve to further prioritise this goal. In order to provide the best possible retrieval and recommendation features to its visitors, MastersPortal needs to invest heavily in “getting to know” them and subsequently providing each and every one of them with a personalised experience.

Although it is out of scope of this thesis to exhaustively explore the full potential of an adaptive recommender system, the results of the analysis that was executed provide MastersPortal with a good sense of what future of its recommender system should be.
The background investigation into recommender systems indicates performance of such a recommender system can be improved by combining a content-based and a collaborative recommender into a single hybrid system.
This fact combined with the initial investigation into an adaptive system makes a clear future direction for MastersPortal apparent: The next iteration of its recommender system should be adaptive, taken into account contextual factors and using both the content-based and collaborative recommenders developed during this thesis. This conclusion is strengthened by the good performance of the current, quite simple, collaborative system employed during the experiments in this thesis.

As such, this thesis provides a stepping stone towards the mid- and long-term enhancements to the MastersPortal.eu website. Focus needs to be placed on introducing more community-based features and the data gathered through these features should be used to build and enhance an adaptive recommender system; utilising both content-based and collaborative approaches.

9. Future Work

The final chapter of this thesis focuses on the future potential of the technologies developed during this thesis and the overall insights gathered. This chapter provides recommendations on future directions that can provide beneficial for MastersPortal based on the results of this thesis.

9.1 Collaborative Recommender

One of the conclusions of this thesis is that currently the content-based recommender slightly outperforms a collaborative recommender. Throughout the thesis this conclusion is heavily nuanced. The core of these nuances is that the collaborative recommender can easily be improved.
An important factor in the functioning of the collaborative recommender is the gathering of good data on visitor preferences. The current approach was implemented purely for the experiments executed as part of this thesis. During the background study and the further analysis of the results many possibilities to improve the process became apparent.

A potential solution mentioned in the background section is Facebook's "Like" system. Visitors use a "like" button to indicate they prefer a certain Master's programme. By aggregating these preferences, and potentially adding "dislikes", a much better collaborative recommender can be build. If this information on preferences is further combined with a visitor profile, the collaborative system certainly has the potential to outperform the content-based approach.
Gathering sufficient data on visitor preference might prove to be much harder than with many of the social networks and major e-commerce sites the ideas in the background chapter stem from. As MastersPortal does not sell directly to its visitors we will need to have other incentives for them to provide us with their preferences. MastersPortal will thus need to invest in compelling community-based features or offer a tight integration with third-parties, such as Facebook.
Other potential directions include close cooperation with university customers in order to gain better insight into what students do once they leave the MastersPortal.eu website. In this respect the challenges faced by MastersPortal are like those faced by the search-engines discussed in *Chapter 2.4.2*.

The information gathered for the collaborative recommender can furthermore benefit the content-based recommender. The information serves as a form of relevance feedback which is used to tune the recommendation based on a visitor's preference.
The enhancements to the collaborative recommender thus make the difference between the collaborative and content-based approaches fade. Automated relevance feedback applied to a content-based recommender based upon visitor profiles effectively turns it into a collaborative recommender. As already noted in the background section, this is an important direction in the current research on recommender systems; they are turning into *hybrid* systems.

9.2 Adaptive Recommender System

The results of this thesis indicate implementing an adaptive recommender system is both feasible on the MastersPortal.eu website and has the potential to improve the recommendations.
As noted in the previous section, much of the data gathered and many of the adoptions necessary to improve the collaborative recommender also serve towards the implementation of an adaptive recommender system. As such implementing an adaptive recommender system on the MastersPortal.eu website seems advisable in the long-term.

The set of contextual factors discussed in this thesis will need to be extended and further refined to enable the adaptive recommender system to perform optimally. The implicit visitor information gathered through the different contextual factors discussed in this thesis, needs to be combined with explicit profile information provided through future community features.
For the MastersPortal.eu website it is imperative to focus on both kinds of data. Using only the explicit information from the community features will leave a lot of users out in the cold. In many cases, such as the situation where a new visitor arrives from Google, it is difficult to match the visitor to existing profiles.

As the adaptive recommender implemented during this thesis focuses purely on feasibility, there are still many question unanswered with respect to the implementation of an adaptive system. Before starting the implementation an adaptive system, the following questions will need to be considered and answered first.
Firstly, on what information do we base our adaptations? Do we use all historical data available, or do we apply some form of rolling horizon? The results of this thesis point towards a rather large variability over time. This variability needs to be understood before any decision can be made.
Secondly, how do we adapt the recommendations? The scoring system as described in this thesis is again only intended the judge feasibility; it is rather simplistic. A better solution is to devise a weighting scheme that generates a single relevancy ranking based on the rankings provided by multiple recommendation approaches. Many different implementations of such a hybrid recommender are already available in literature.
Finally, how do we track performance of the system after applying the adaptations? A potential solution is continuous testing, in which a small portion of the visitors is used to tune the system for remainder of the visitors. Such an approach requires a very careful setup as otherwise biases might be introduced. Other, less dynamic, options are also possible. We could for example run an experiment every once in a while to gather the new settings for the adaptive system.

Before MastersPortal turns its attention to an adaptive recommender system, it seems best to first focus on implementing community-based features. Subsequently, a hybrid recommender system utilising information from these features can constructed.
Once such a system is up and running, adaptive features can be introduced. These should initially focus on optimising the performance of the hybrid recommender, but later on can also be applied in other parts of the MastersPortal.eu website.
A preliminary literature review on adaptive recommender systems, as presented in *Chapter 2.3*, shows that much research is currently undertaken in this field. Before considering actually implementation an adaptive system, the current state-of-art should be re-evaluated.

9.3 Experimentation Framework

The experiments executed as part of this thesis provide the basis for a future experimentation framework within MastersPortal. There is still a lot of work that will go into implementing a fully functional experimentation framework.
As noted in the background chapter at the start of this thesis, Google provides a system which can be implemented free of charge through its *Website Optimiser* service. There are undoubtedly other initiatives which provide the same. MastersPortal will need to investigate these initiatives and decide whether they can be combined with the current basis or whether more custom development is required.

Apart from the technical implementation of an experimentation framework, MastersPortal will also need to think carefully of the kinds of questions that it wants answered.
Some questions are concerned with lay-out issues; the placement of features such that they attract the most attention. Similar issues arise concerning the best placement of banner advertisements. All these questions are directly related revenue generated through the MastersPortal.eu website.
Questions with a less direct effect on revenue are also available in ample supply. An important point, shortly addressed in the previous section, is the optimisation of MastersPortal's search-engine. The current user interface to this search-engine offers a lot of options without knowing if these options are of value. Providing the best possible user experience is an important part of MastersPortal's future success.
Both types of questions are valid, but both need to be treated differently with respect to the future experimentation culture within MastersPortal. Questions strongly linked to an immediate increase in revenue are often easily measured and relevant in the short term.
The indirect revenue questions are more open to interpretation, but offer the possibility to provide structural improvements to both the website and its underlying concepts. Striking a good balance between both types of questions is of paramount importance for an experimentation culture within MastersPortal to provide meaningful results.

References

Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. *ACM Transactions on Information Systems* , 103-145.

Balabanovic, M. (1997). An adaptive Web page recommendation service. *Proceedings of the first international conference on Autonomous agents* (pp. 378-385). Marina del Rey: ACM.

Balabanovic, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM* , 66-72.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence.* Madison: Morgan Kaufmann Publisher.

Cooper, D. R., & Schindler, P. S. (2003). *Business Research Methods.* New York: McGraw-Hill.

Crook, T., Frasca, B., Kohavi, R., & Longbotham, R. (2009). Seven Pitfalls to Avoid when Running Controlled Experiments on the Web. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1105-1114). Paris: ACM.

Dey, A. K. (2001). Understanding and Using Context. *Personal and Ubiquitous Computing* , 4-7.

Domingues, M. A., Jorge, A. M., & Soares, C. (2009). Using Contextual Information as Virtual Items on Top-N Recommender Systems. *Workshop on Context-Aware Recommender Systems (CARS-2009).* New York: ACM.

Garcia, E. (2006, October 26). *Cosine Similarity and Term Weight Tutorial*. Retrieved October 27, 2009, from Mi Islita: http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html

Joachims, T. (2002). Optimizing Search Engines using Clickthrough Data. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 133-142). Edmonton: ACM.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2008). Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* , 140-181.

Lam, W., Ruiz, M., & Srinivasan, P. (1999). Automatic Text Categorization and Its Application to Text Retrieval. *IEEE Transactions on Knowledge and Data Engineering* , 865-879.

Lam-Adesina, A. M., & Jones, G. J. (2001). Applying Summarization Techniques for Term Selection in Relevance Feedback. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 1-9). New Orleans: ACM.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

MaxMind. (2009, December 1). *GeoLite Country*. Retrieved December 17, 2009, from MaxMind.com: http://www.maxmind.com/app/geolitecountry/

McClave, J. T., Benson, P. G., & Sinicich, T. (2001). *Statistics for Business and Economics.* Upper Saddle River: Prentice Hall.

McKirnan, D. J. (2010, January 21). *Basics of Scientific Method.* Retrieved May 1, 2010, from Psychology Research Methods: http://www.uic.edu/classes/psych/psych242/Week2.html

Montgomery, D. C., & Runger, G. C. (2003). *Applied Statistics and Probability for Engineers.* New York: John Wiley & Sons.

Nunamaker, W., Chen, M., & Purdin, T. (1990). Systems development in information systems research. *Journal of Management Information Systems* , 89-106.

O'Mahony, M., Hurley, N., Kushmerick, N., & Silvestre, G. (2004). Collaborative recommendation: A robustness analysis. *ACM Transactions on Internet Technology* , 344-377.

Pazzani, M. J., & Billsus, D. (2007). Content-Based Recommendation Systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl, *The Adaptive Web* (pp. 325-341). Berlin Heidelberg : Springer-Verlag.

Porter, M. (2006, January). *The Porter Stemming Algorithm*. Retrieved October 25, 2009, from Tartarus: http://tartarus.org/~martin/PorterStemmer/

Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting clicks: estimating the click-through rate for new ads. *Proceedings of the 16th international conference on World Wide Web* (pp. 521-530). Banff: ACM.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gull, A., & Lau, M. (1995). Okapi at TREC-3. *The Third Text REtrieval Conference (TREC-3).* NIST.

Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender Systems in E-Commerce. *Proceedings of the 1st ACM conference on Electronic commerce* (pp. 158-166). Denver: ACM.

Sculley, D., Malkin, R. G., Basu, S., & Bayardo, R. J. (2009). Predicting Bounce Rates in Sponsored Search Advertisements. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1325-1334). Paris: ACM.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *SIGIR '96* (pp. 21-29). ACM.

United Nations Statistics Division. (2009, October 1). *United Nations Statistics Division- Standard Country and Area Codes Classifications (M49)*. Retrieved February 22, 2010, from United Nations: http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm

Welch, B. L. (1947). The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika* , 28-35.

Yang, Y., & Chute, C. G. (1994). An Example-Based Mapping Method for Text Categorization and Retrieval. *ACM Transactions on Information Systems* , 252-277.

Appendixes

A. Visitor Statistics – October 2009

Four weeks, Monday through Sunday, from October 5th up to and including November 1st of 2009 are analysed in more detail to provide information on the bounce-rate and revisit-rate for visitors to the MastersPortal.eu website.
Furthermore, these results, generated by a custom log-file harvester, are compared against the results generated by a proven third-party product to ensure their validity.

The statistics presented in this appendix are harvested directly from the log-files of the Apache2 web server serving the MastersPortal.eu website. For this purpose, a custom log-file harvester, the "MastersPortal Harvester" was constructed. In order to verify the correctness of this custom harvester, the results presented in this appendix are first compared against the results provided by the Webalizer log-file harvester. Webalizer is a trusted, third-party, log-file harvester in use within MastersPortal as its main source of information on visitor statistics.

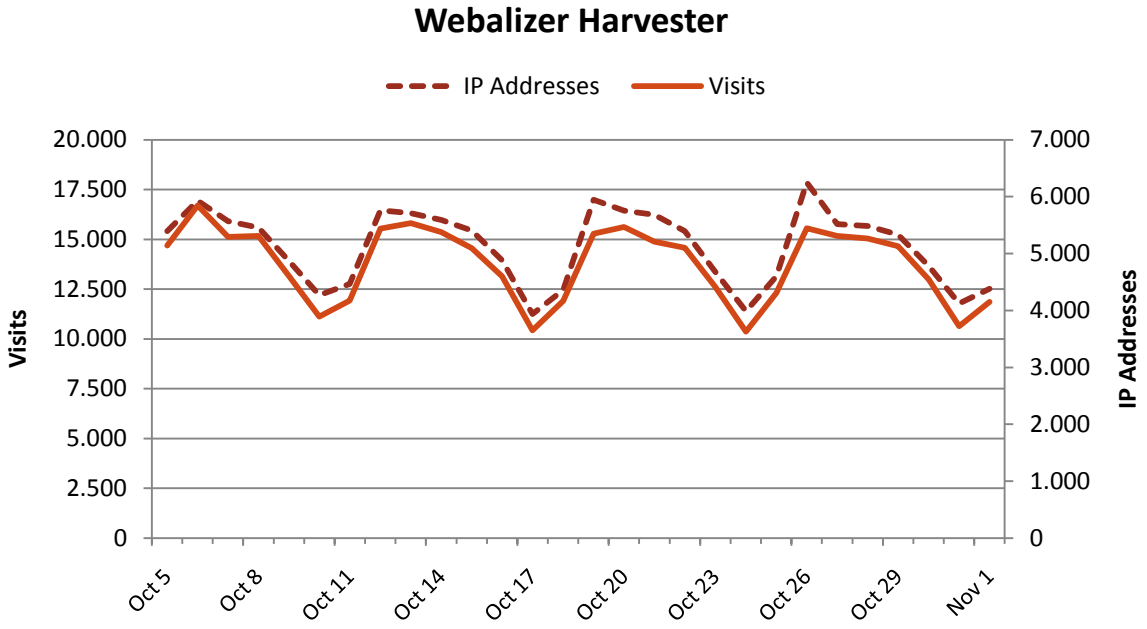


Figure 36: Daily number of visits and unique IP addresses, according to the third-party Webalizer harvester

In *Figure 36* above, the results of the Webalizer harvester are displayed. A clear weekly trend is visible from these results. Generally speaking, Monday and Tuesday are the busiest; Saturday has the lowest number of visitors. Taking a month-over-month perspective there is a clear growth in the total number of visitors, which is only faintly visible in the week-over-week view presented by the figures in this appendix.

The graph in *Figure 36* allows for a relative comparison of the number of visits and unique IP addresses. As is clear from this graph, there is a strong correlation between the number of unique IP addresses and the number of visits in the four weeks measured. Note that the daily sum of unique IP addresses is higher than the total number of unique IP addresses recorded for the four week period. The same unique IP address can be counted on multiple days.

Below, in *Figure 37*, the same results are presented, but this time generated by the custom log-file harvester created specifically for this thesis.

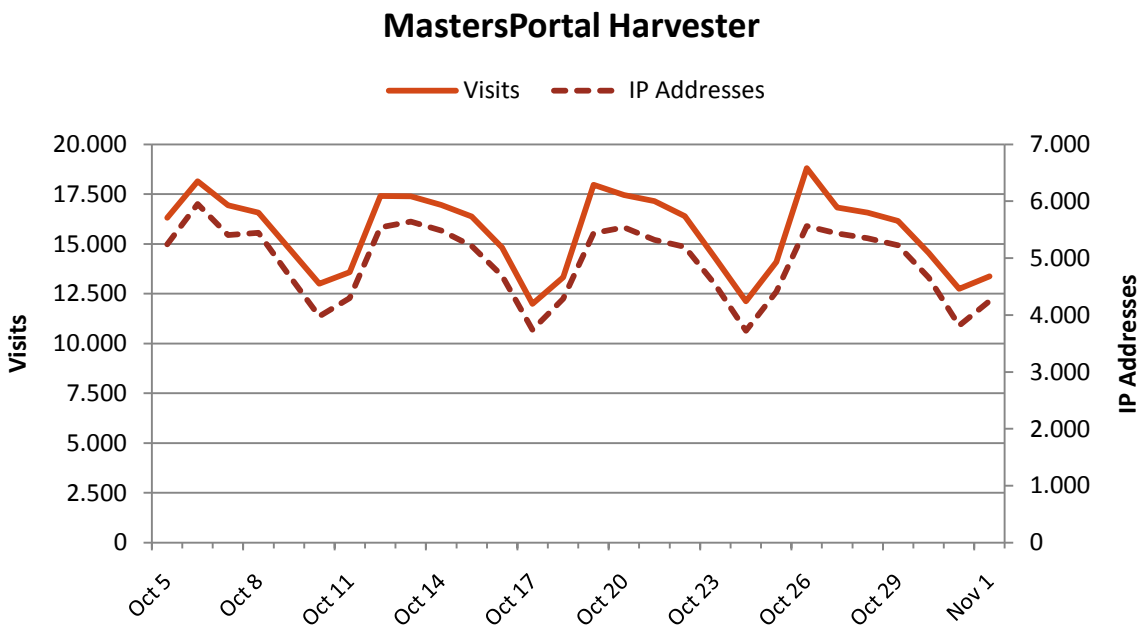


Figure 37: Daily number of visits and unique IP addresses, according to the MastersPortal harvester

We again see the same strong correlation between the number of visits and the number of unique IP addresses. In the remainder of this appendix I will therefore use the number of unique IP addresses to verify the correlation with other metrics.

An initial, visual, comparison of the IP address numbers provided by both Webalizer and the custom log-file harvester indicate both datasets are in line.

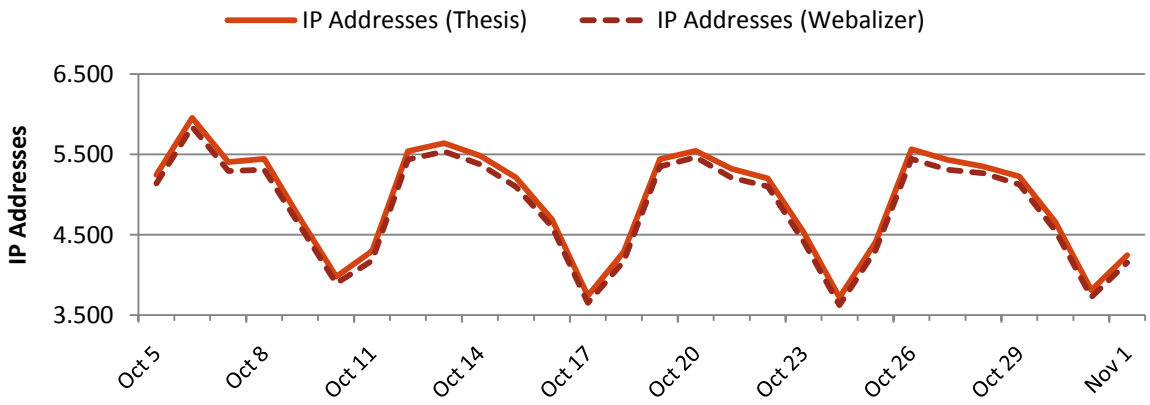


Figure 38: Daily number of unique IP addresses reported by the MastersPortal Harvester (Thesis) and Webalizer

In *Figure 38* both sets of IP address numbers are graphed together, which confirms the initial visual diagnosis: Both the Webalizer harvester and the custom log-file harvester constructed for this thesis provide similar results.
The small difference between the two harvesters is caused by differences in the filtering of potential non-human visitors.

The custom log-file harvester recorded visitors from 98.205 unique IP addresses during the four weeks analysed, resulting in 436.042 visits to the MastersPortal website.
Of these visits, 22.042 came directly to the MastersPortal.eu homepage. 151.438 Came into one of the, at that time, approximately 12.000 pages offering detailed information on a single Master's programme.

A1. Bounce-Rate

Having verified the proper operation of the custom log-file harvester, its results were used to determine the bounce-rate for visitors to the MastersPortal.eu website over the course of a four week period running from October 5th until November 1st of 2009.

As already made clear in the introduction to the thesis, the single largest external referrer to the MastersPortal.eu website is Google's search-engine. Google's search-engine is thus the primary subject of this analysis. In order to provide a frame of reference, a second analysis is performed, calculating the bounce-rate for the combination of all other, not search-engine, external referrers.

The results for Google's search-engine are presented in *Figure 39*. In this figure visitors from other Google websites, such as Gmail and Google Translate are explicitly excluded.

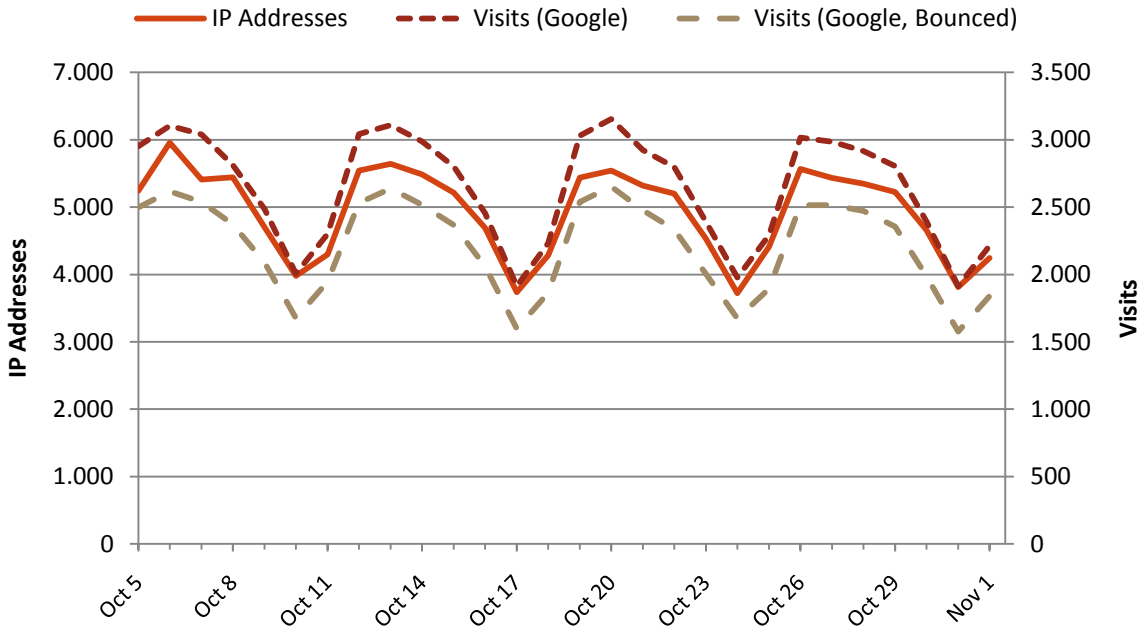


Figure 39: Number of unique IP addresses mapped against visits, total and bounced, from Google's search-engine

Again, the very distinctive weekly pattern is clearly visible. In the graph, the number of unique IP addresses is mapped on the left vertical axis; the number of visits is mapped on the right vertical axis. This allows for a relative comparison of both datasets.

The comparison shows a strong correlation between the number of IP addresses, visits and bounced visits. The bounce-rate, calculated as the percentage of total visits that has bounced, for visitors coming from Google is steady at nearly 84% (= 83,88%; σ = 0,70%). A graphical overview of this bounce-rate is provided in *Figure 4* in the introduction of this thesis.

The next graph, in *Figure 40*, again shows the total number of IP addresses but this time graphs them against the number of other, not search-engine, visits and bounced visits. Although somewhat less

obvious, the relation between the number of IP addresses and the number visits is still present. The relation between the number of visits and bounced visits is just as strong as in the previous graph. The instability in the number of visits is caused by the far fewer number of "other" referrals the MastersPortal.eu website receives. Whereas Google provides around 2.600 visits on a daily basis, only on average 70 other referrals are counted.

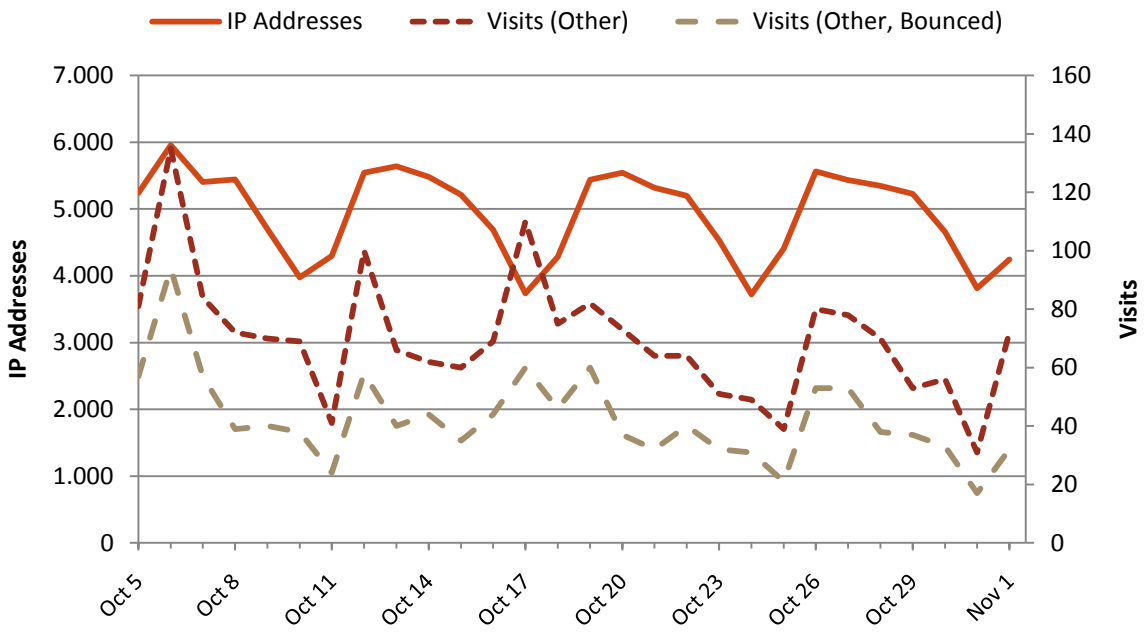


Figure 40: Number of unique IP addresses mapped against visits, total and bounced, from "other" referrers

The bounce-rate for other referrals is less stable around 65% ($\bar{x} = 64{,}47\%$; $\sigma = 7{,}26\%$), but it is clearly well below the average Google bounce-rate. Both bounce-rates are again visualised in *Figure 4* in the introduction of this thesis.

A2. Revisit-Rate

Apart from the bounce-rate discussed in the previous section, a second, related metric is of importance too. This metric is the revisit-rate. The revisit-rate is defined as the number of times a visitor comes back to the MastersPortal.eu website. This analysis is executed by looking at the number of visits (sets of requests with at most sixty minutes in between them) a single IP does to the MastersPortal.eu website from October 5th up to and including November 1st.
For this analysis we consider only visitors from Google's search-engine. The "other" category discussed in the previous section is not taken into account.

Out of 74.043 visits coming in from Google, 55.337 distinct IP addresses are counted. From these distinct IP addresses only 10.094 are counted as having more than one visit. Thus, once bounced, only $18{,}24\%$ of visitors come back for a second visit at a later point in time.
This result underlines the conclusion that visitors, referred to the MastersPortal.eu website through Google, need to be convinced to stay on the MastersPortal.eu website as quickly as possible, preferably on the first page they view.

The graph in *Figure 41* provides an additional illustration of the conclusion drawn above. It shows an overview of the top 200 visitors coming from Google, looking at the number of times they were

referred. This metric is strongly related to the revisit-rate, as visitors with only a single referral also have only a single visit.
The goal of this graph is to show that within the approximately 10.000 visitors who visit multiple times, only a relatively small numbers of visitors visits a substantial number of times. It is an indication of how many chances the MastersPortal.eu website gets to "convert" the average visitor coming in from Google.

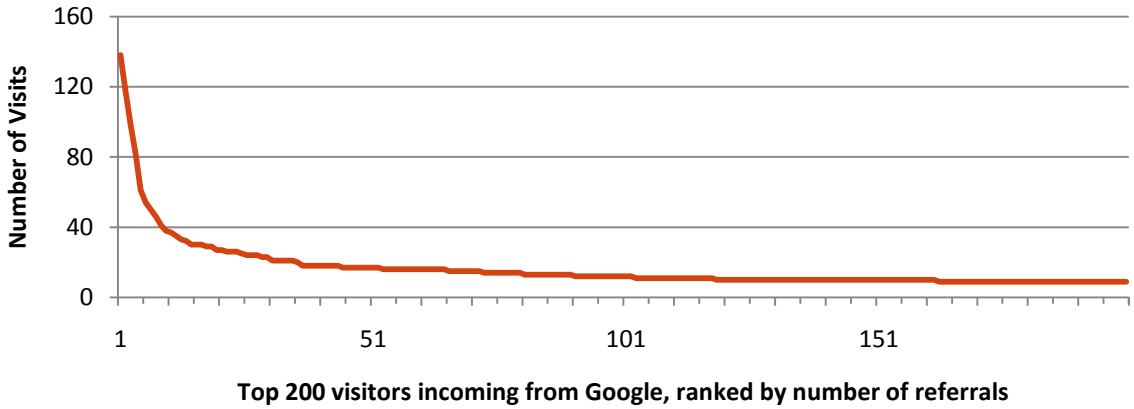


Figure 41: Number referrals from Google for the top 200 visitors to MastersPortal.eu

In *Figure 41* we see a very long-tailed distribution. Only a small portion of the total number of visitors referred to the MastersPortal.eu website by Google visits multiple times. Apart from a few initial exceptions, all visitors in the top 200 visit less than twenty times. Combining this result with the revisit-rate computed earlier strengthens our conclusion: the MastersPortal.eu website needs to convert visitors coming in from Google as quickly as possible.

B. Experiment Implementation Details

This appendix provides details on the implementation of both experiments executed during of this thesis. The goal of this appendix is to provide a high-level overview. With the details provided in this appendix it is possible to produce similarly functioning experiments.

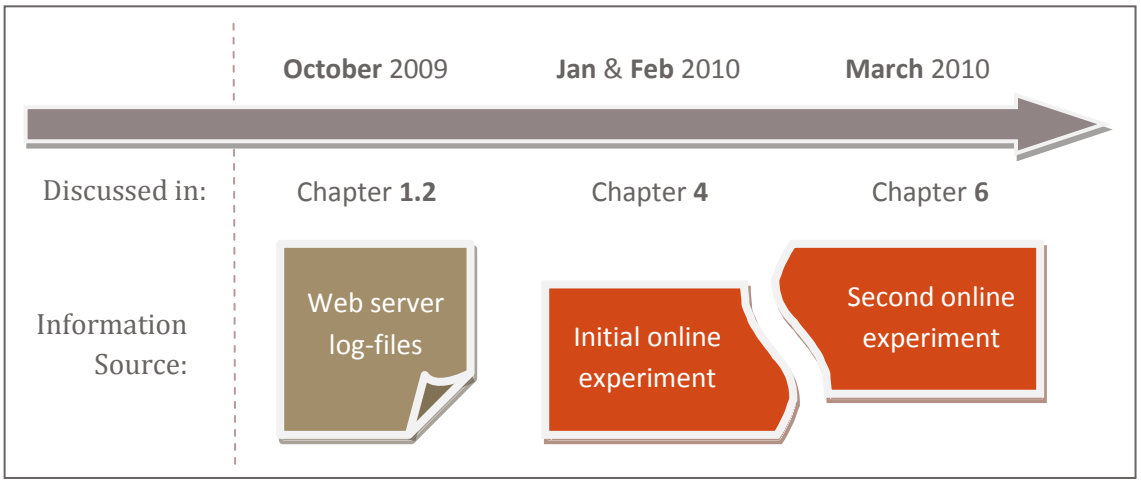


Figure 42: Timeline of the data gathering steps

Before the overall architecture of the experiments is discussed, *Figure 42* above provides a timeline of the data gathering steps at the heart of the experiments executed during this thesis. The diagram provides a chronological background to the subsequent discussion.

B1. Overall Architecture

The MastersPortal.eu website is powered by a custom-build *Content Management System*. This system is written in object-oriented PHP 5.3[11] and runs on top of the Apache 2.2[12] web server. It utilises a MySQL 5.0[13] database as its primary data store.

The exact implementation of the MastersPortal *CMS* is not relevant to this discussion. It is though relevant to note that the CMS is a modular system; to execute the experiments a single module was added to the system. This module was placed on all web pages containing detailed information on the Master's programmes in the MastersPortal database.

A second important design feature of the CMS is the fact that it is *stateful*. All visitors are tracked throughout their visit by means of a session cookie and a server-side session data store. Each request send by a visitor is thus fully aware of the entire state of the visitor's session and, previous, requests send by the visitor.

For all recommender systems deployed during the experiments, the top 20 recommendations for each programme were pre-computed and stored in the MySQL database. During the experiment, retrieving any recommendation was thus nearly instantaneous; the computational performance of the recommender approaches was completely factored out.

The implementation of both the content-based and collaborative recommenders is discussed in more detail in *Appendix D*.

[11] http://www.php.net/
[12] http://httpd.apache.org/
[13] http://dev.mysql.com/

All data gathered during the experiments was stored in the MySQL database. The analyses presented throughout this thesis are executed by either querying the database directly or through the means of intermediate export scripts.
The exports scripts retrieve data from the MySQL database and store it in a format readable by Microsoft Excel. All further analyses are executed using Excel. Especially Excel's pivot table functionality proved to be a versatile solution to interpreting and visualising the results of the experiments. Final conclusions on statistical significance are based on analyses executed using StatGraphics Centurion[14].

B2. Participant Tracking and Logging

This section describes how participants were tracked throughout the experiment and how the participant tracking information was stored.
Upon their first arrival to a programme details page on the MastersPortal.eu website each visitor is assigned an *Experiment ID*; a unique 64 character hexadecimal string. This is string is stored in a permanent cookie within the visitor's browser.
This string is used to track the visitor throughout the experiment. Over the course of the experiment all details concerning the visitor's actions are linked to the participant referenced by this unique experiment ID. The information stored for each participant is provided in *Table 24*.

Value	Type	Description
Participant ID	Integer	Unique database ID
Previous ID	Null *or* Integer	ID of suspected previous participant
Experiment ID	String	Unique identification string
IP	Integer	IP address
Country	String	ISO country-code
Group	Enumeration	Experiment group assigned to participant
Session Count	Integer	Number of sessions started
Cookie Error	Boolean	Set to true when cookie issues are detected
Created	Integer	Timestamp of creation
Updated	Integer	Timestamp of last update

Table 24: Information stored for each participant during the first experiment.

Whenever a cookie-error is detected for a participant, the conflicting *Participant ID* is stored as the *Previous ID*. This links the newly created participant to its suspected predecessor. By following the chain thus created it is possible to merge the "cookie-error participants" back into a single entity. Due to cookie-errors causing violations within the controlled experiment, this feature is not further exploited as part of this thesis. A further discussion on the detection of cookie-errors is provided below.
During the second experiment some additional information was stored for each participant; an overview is provided in *Table 25*.

Value	Type	Description
Continent	String	ISO continent-code
Resolution	Enumeration	Screen Resolution context (see *Chapter 5.2.4*)

Table 25: Additional information stored for each participant during the second experiment.

[14] http://www.statgraphics.com/

Each recommendation received by the participant is stored in the database as well. The information stored for each recommendation is provided in *Table 26*.

Value	Type	Description
Recommendation ID	Integer	Unique database ID
Programme ID	Integer	Programme the recommendation is for
Previous ID	Null or Integer	ID of previous recommendation
Referrer	String	Referrer
Time	Integer	Timestamp of recommendation
Programme Count	Integer	Number of programmes recommended
Session Count	Integer	Session the recommendation is issued in

Table 26: Information stored for each recommendation during the first experiment.

Again some additional information was stored during the second experiment. An overview of this information is provided in *Table 27*.

Value	Type	Description
Query	Enumeration	Google Query context (see *Chapter 5.3.3*)
Discipline	Enumeration	Academic Discipline context (see *Chapter 5.2.3*)

Table 27: Additional information stored for each recommendation during the second experiment.

Finally, whenever a recommendation is clicked, details of the "click" are also stored. An overview of the information stored is provided in *Table 28*. Note that the details of the click are stored upon loading the page where the "click" points to. They are not stored when the actual click occurs.

Value	Type	Description
Programme ID	Integer	Programme the clicked item recommends
Position	Integer	Position of the clicked recommendation (ranging from 1 to 8)
Time	Integer	Timestamp of click

Table 28: Information stored for each click.

To properly track clicks generated by the experiments' participants, and to be able to separate them from regular clicks, a tracking-code is attached to each link presented by the experimental programme recommender. An example tracking-code is provided in *Figure 43*.
The tracking-code does not provide any new information; it merely serves as a trigger to link existing pieces of information together. As such, the tracking-code inherently ensures its own integrity. If tampered with, the tracking-code generates a mismatch with information already stored.

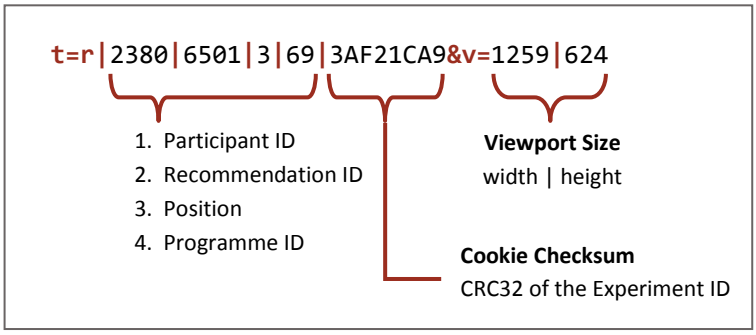


Figure 43: Example decomposition of the experiment tracking-code

The final part of the tracking-code consists of the viewport size of the visitor. It is appended through the use of client-side JavaScript code. As some visitors disable JavaScript support and some browsers simply failed at detecting the viewport size this part is not always included.

Upon loading a programme details page the tracking-code, if present, is analysed. All values in the tracking-code need to match values already in the system: The *Participant ID* in the tracking-code should match the one stored in the CMS session; the *Programme ID* should match the programme currently displayed; the programme currently displayed should be present in the provided *Recommendation ID* at the provided *Position*. If a mismatch occurs within any of these values, the click is ignored. Duplicate clicks are also ignored.
In order to ensure cookie validity, the *Experiment ID* is also included in the tracking-code. To keep the length of the tracking-code limited, an eight character hexadecimal checksum is used. If the checksum does not match the *Experiment ID* stored in the participant's cookie, the participant is assumed to not have properly handled his experiment cookie. These participants are excluded from all further analyses. In *Appendix F* these participants are listed as part of the "not accepting cookies" classification. Cookie-errors of this sort were detected in less than 1% of all participants.

Apart from storing information in the database, a backup plain-text log-file is also maintained during the experiment. An overview of the information stored in this plain-text log-file is provided in *Table 29*. The plaint-text log-file includes some information that is not stored in the database. This information is deemed interesting but not immediately useful. It is stored just in case a future analysis would require it. The analysis presented in *Chapter 5* makes use of some of this additional information.

Value	Type	Description
Time	Integer	Timestamp of recommendation
IP	Integer	IP address
Participant ID	Integer	Database ID of participant
Recommendation ID	Integer	Database ID of recommendation
Programme ID	Integer	Database ID of programme
Group	Enumeration	Experiment group assigned to participant
Viewport	String	Width and height of the participant's viewport
History	String	List of previous recommendations for the participant
Referrer	String	URL of the referring page
User Agent	String	User agent string

Table 29: Information stored for each recommendation in the plain-text log-file

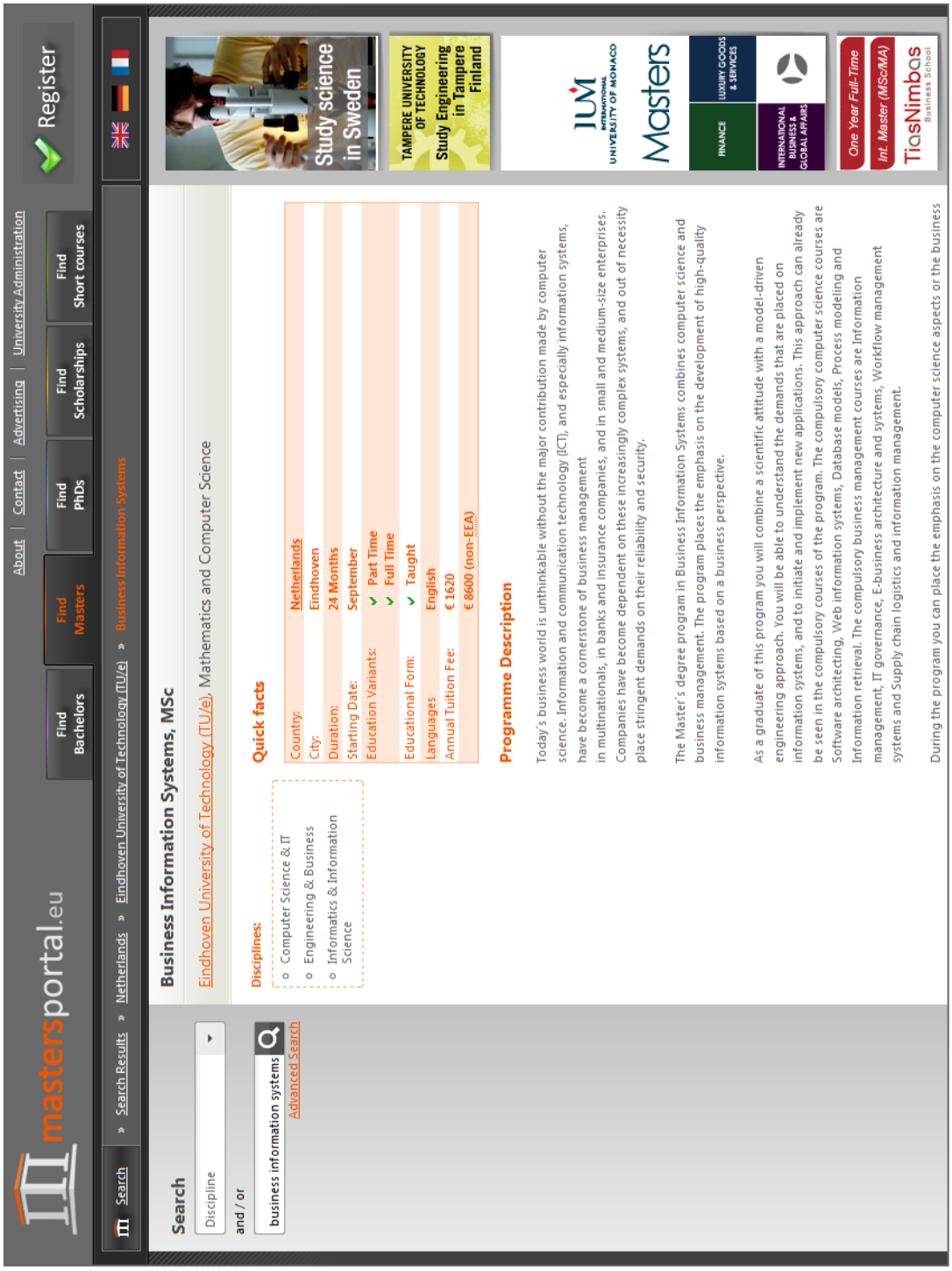
Finally, as the tracking-codes are self-contained and appended to the URL's requested by the visitors, the web-server log-files contain a third copy of the experiment's data. The full validity of this data is though difficult to ascertain without resorting to either one of the two primary data sources maintained during the experiment.

C. Screenshots of the Programme Recommender

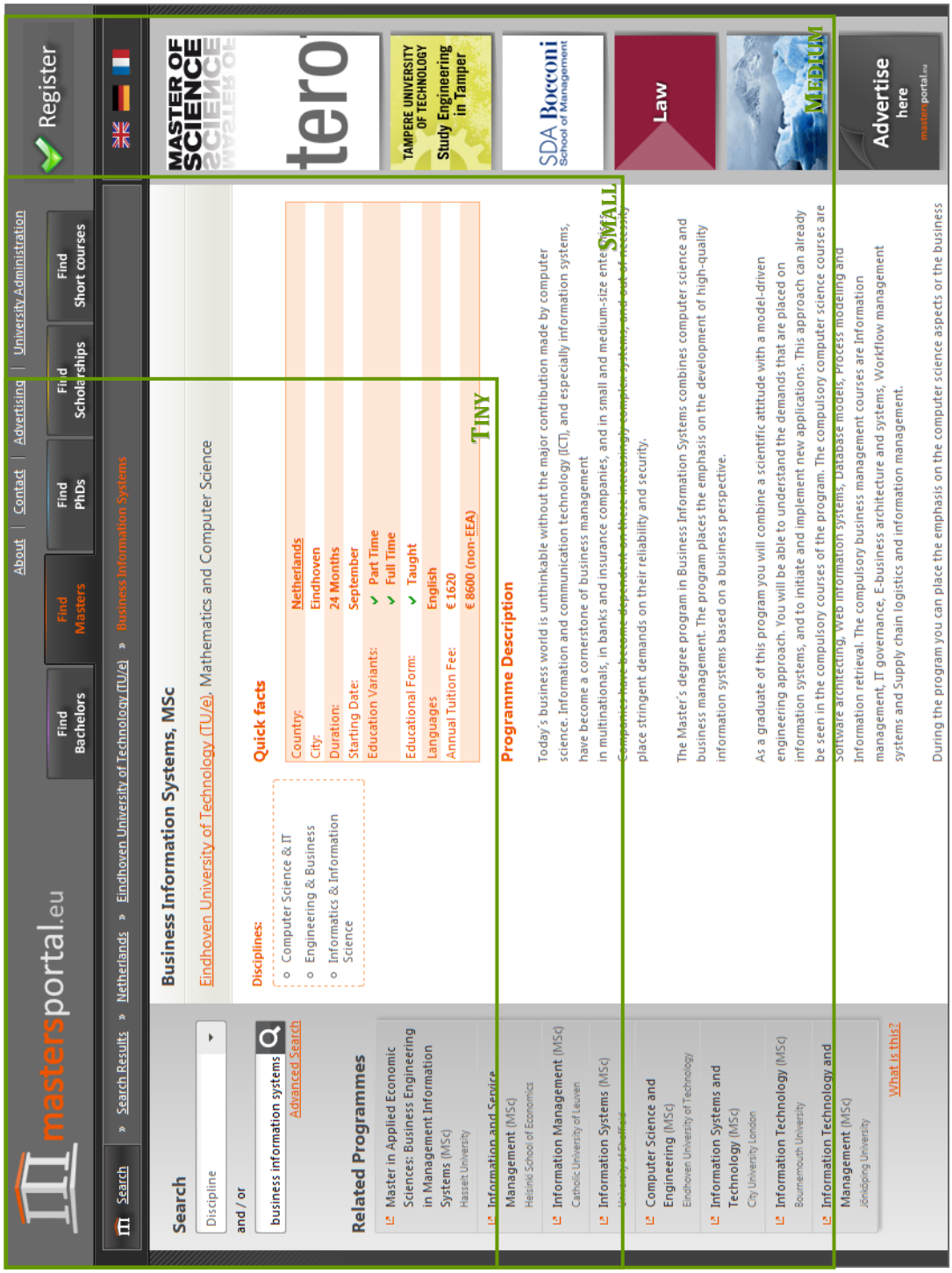
This appendix provides screenshots of the programme detail pages on the MastersPortal.eu website during the first experiment. It shows how the layout of the page differed for participants in the "*Control – None*" group versus participants in the other experiment groups.

In addition, the screen resolution categories as discussed in *Chapter 5.2.4* are overlaid on the second screenshot. This allows for a better understanding of how this categorisation came about.

C1. Programme Details Page without Recommender



C2. Programme Details Page with Recommender



D. Recommender Implementation Details

This appendix provides additional details on the implementation of the two automated and more advanced recommender approaches discussed in *Chapter 3*. It provides an overview of the assumptions made while implementing the recommenders and its lists any other issues that were encountered during their implementation.

D1. Content-Based Recommender

The recommendations computed by the content-based recommender are generated through a tf-idf based vector-space comparison of the programme descriptions against each other. The exact implementation of the tf-idf algorithm is described the background of this thesis, *Chapter 2.2*.

Due to the amount of programmes in the MastersPortal database, the computational power required to process each recommendation is too large to implement a real-time recommender. The recommendations were thus pre-computed.
Generating a complete recommendation (c.q. ranking all programmes in relevance based upon the reference programme) for a single programme takes on average 23 seconds on my laptop. Pre-computing all recommendations will thus take around 72 hours, which makes pre-computing up to this extent also infeasible.

I therefore had to look into ways to reduce the time required to compute the recommendations. An initial observation is that reducing the number of terms to be compared exponentially reduces the time required to compute a recommendation.
A second observation is that the complete recommendations, as mentioned above, are highly susceptible to noise. For example, programmes concerning “logistics performance” get a lot of highly ranking recommendations in the field of “music performance”, due to the overlapping concept of “performance”.
A similar problem is observed by (Lam-Adesina & Jones, 2001) who evaluate the performance of their query expansion technique. The authors aim to prevent unrelated terms from showing in their expanded query, effectively what happens in the example in the previous paragraph. Their technique is to not use the entire text of a document, but only an automatically generated summary as the basis for query expansion. According to the authors, query expansion using the automated summary improves its performance.
Finally, an important factor in the ranking generated for the MastersPortal programme recommender is that we do not require the best possible ranking. For the MastersPortal.eu website only the top of the relevancy ranking is required. This top ranking is used to recommend a limited set of programmes to the visitor. As such, a close approximation of the best possible ranking will suffice.

As a result, a recommender is implemented that only takes into account the 50% most relevant terms for each programme based on their *tf-idf* values. This is effectively the same as automatically generating a summary as proposed by (Lam-Adesina & Jones, 2001).
The computing time required to process a recommendation decreased nearly ten-fold this way. Using this approach, it is now feasible to pre-compute all required recommendations overnight.

D2. Collaborative Recommender

For the collaborative recommender, visitor-behaviour is extracted from the web server log-files. This process utilises a procedure similar to the procedure used to compile the initial statics presented in *Chapter 1.2* of this thesis.
For the first experiment, log-files starting on the 10th September 2009 up to and including the 10th of January 2010 were analysed.

From these log-files, 266.475 programme recommendations based on 24.322 distinct visitor sessions were harvested. On the 10th of January, the MastersPortal database contained approximately 14.000 Master's programmes.

Relevant programme recommendations originate from visitor sessions that:

- Are not generated by non-human visitors;
- Contain visits to at least 4 different programmes;
- Contain visits to at most 36 different programmes;
- Have a *Tanimoto coefficient* of below 0,9 compared to any of the other groups harvested;

These definitions are set-up to be as conservative as possible, erring on the side of caution and thus potentially excluding relevant information. Since a large amount of visitor-behaviour is available, it is in my opinion better to exclude some relevant information than to potentially include irrelevant information.

As a result of this filtering, the following information was excluded:

1.039.776 Non-human generated visitor sessions:

These are visits generated by the crawlers of for example Google and Yahoo! They were identified by comparing the "user-agent" string provided by the client against a list of crawlers known to visit the MastersPortal.eu website.

52.720 Sessions of visitors viewing 1, 2 or 3 programmes:

"Groups" with a size of one programme are excluded as they are of no use to the recommending process. Groups with a size of two or three programmes are excluded because they might constitute accidental groupings. These are generated by visitors who are not actually searching the MastersPortal database for contextually similar programmes, but just happen to view more than a single programme.

526 Visitor sessions containing more than 36 programmes:

Although non-human visitors are filtered through their user-agent string, not every non-human visit is caught this way. Some non-human sessions are still present in the dataset, showing up as outliers.

The most obvious outliers are visitors who, in a single session, view between 10% and 90% of *all* programmes in the MastersPortal database. Groupings generated by these visitors can clearly be considered irrelevant.
Visitor sessions that contain more than 36 programmes are considered to be outliers. This number is dynamically generated during processing and set to be equal to three times the average number of programmes in a group. Since it is difficult to quantitatively distinguish a valid visit from a visits generated by a non-human visitor I have again opted to take a conservative approach.

The limit of three times the average number of programmes is well below the point where I expect most non-human visitors to be. At the time of writing this thesis, this suspicion is confirmed by *Table 33*, which lists the results of behavioural identification of non-human visitors during the first experiment.

46 Groups were removed because of similarity constraints:

A group is considered to be the duplicate of another group if its Tanimoto coefficient (*Jaccard coefficient* in case of binary attributes) is less than $0{,}10$. In other words: More than 90% similar to any other group.

Although it is very well possible that several users generate similar groups, the change they generate exactly the same group is very slim, especially when taking into account that the small groups have already been removed. The result of this filtering step also confirms this assumption. Only 46 additional groups were removed.

E. Alternative Multiple Comparison of Means

As discussed in *Chapter 4.4*, the analysis of the results of the first experiment contains an important assumption about the sizes of the different experiment groups: They are assumed to equal. This is in reality not the case. The intention of this appendix is to show that even though this assumption influences results slightly it has no effect on the overall conclusions presented in the chapter.
Note that this alternative analysis method does *not* affect the actual bounce-rates as computed during the first experiment. It merely affects the statistical significance of the difference between the bounce-rates of the different groups.

For the analysis presented in this appendix a multiple comparison procedure was again applied using the Bonferroni method and a 99% confidence interval. The only difference is that this analysis is executed with the actual group sizes instead of assumed equal group sizes.

In *Table 30* below, the group overview constructed from the multiple comparison procedure is displayed. The table looks quite similar to *Table 2* in *Chapter 4.4*, with the exception of the content-based and collaborative recommenders; they are now both considered heterogeneous groups.

	Homogeneous Groups
None	•
Random	•
Baseline	•
Collaborative	•
Content	•

Table 30: Homogeneous groups within the alternative multiple comparison procedure

In *Table 31* the full contrasts as computed through the multiple comparison procedure are displayed, just as previously in *Table 4*. Two columns are added to the table: The difference in number of participants between the groups, n, and the difference in bounced visits, μ.
The final difference, expressed in bounced visits, is determined by subtracting μ from n. Where this difference is outside of the listed confidence interval, the contrast between the groups is considered to be significant; the group performance is not equal.

Contrast	Difference (n)	Difference (μ)	Difference		Confidence
Baseline – None	-753,0	-1252,0	-499,0	*	60,15
Baseline – Random	-2,0	-550,0	-548,0	*	61,07
Baseline – Content	-42,0	431,0	473,0	*	61,02
Baseline – Collaborative	879,0	1122,0	243,0	*	62,29
None – Random	751,0	702,0	-49,0		60,14
None – Content	711,0	1683,0	972,0	*	60,09
None – Collaborative	1632,0	2374,0	742,0	*	61,38
Random – Content	-40,0	981,0	1021,0	*	61,01
Random – Collaborative	881,0	1672,0	791,0	*	62,29
Content – Collaborative	921,0	691,0	-230,0	*	62,23

Table 31: Group contrasts for the alternative multiple comparisons (* indicates statistically significant difference)

The major difference for the alternative analysis is the fact that the content-based and collaborative recommenders are not considered to be homogeneously anymore, by a substantial margin. When we order the contrasts by their absolute size we still see the same pattern: The "Content – Collaborative" contrast is apart from the "None – Random" contrast the smallest.

From the above analysis we thus need to conclude that the difference in bounce-rate between content-based and collaborative recommenders is significant. As already noted in *Chapter 4.4*, the approach used in the main analysis points the same conclusion. It indicates the difference is just barely outside of the confidence interval.

My major reservation with the approach used in this appendix is the interpretation of contrasts as mentioned in *Table 31*. The contrasts are calculated by simply subtracting the absolute difference between the groups from the difference in bounced visits. This does not take into account the fact that if both group sizes were equal the relative number of bounced visits would also be different between the groups, because of their inherently different bounce-rates.
This is better taken into account in the main analysis. The truth, for as far as it is possible to speak of this, is most likely in between the results presented here and those presented in the main analysis.

Either way, both analyses support the overall conclusions drawn in *Chapter 3*. As the ambiguity of the difference between the content-based and collaborative recommenders is discussed in detail throughout this thesis, the results from this appendix have no additional consequence.

F. Data Cleanup and Filtering

As part of the analysis of the results of both experiments, extensive data cleanup and filtering is applied. This is done to get an as realistic result as possible. The goal of filtering is to remove all non-human visitors and to remove all visitors who violate the conditions of the controlled experiment.

The cleanup process consists of three steps: Firstly there is a pre-filtering step which is applied during the experiment itself. This step filters based upon the user-agent string reported by most visitors. Its goal is to remove all non-human visitors who behaved "nicely". They identify themselves as a non-human visitor and can thus be easily filtered.
During both experiments the same pre-filtering was applied. In *Table 32* below the user-agents filtered are listed. If any of the strings in the table below are present in the user-agent string presented by the visitor, this visitor has been pre-filtered from the experiment.

Non-Human User-Agents

Slurp	facebookexternalhit
Googlebot	Yeti
Mediapartners-Google	WebAlta Crawler
msnbot	ia_archiver
Baiduspider	INGRID
Charlotte	Gigabot
Ask Jeeves	SapphireWebCrawler
Twiceler	NEWT ActiveX
DotBot	Yandex

Table 32: Non-human user-agents

The other two filtering steps are applied after the experiment is completed, but before the actual analyses are executed. These two steps are an IP filter, which removes visitors at IP address level and a participant filter which removes participants from the experiment.

The IP filter aims to remove all non-human visitors who do not identify themselves properly. It furthermore removes all visitors who do not accept the experimentation cookie.
This filtering step is based upon visitor-behaviour. Non-human visitors are identified by the large number of recommendations they generate over a short period of time. Visitors not accepting the experimentation cookie are identified by the fact that each request they do results in a new participant being created. This leads to a large number of participants from the same IP address in a short time span.

The final filtering step is a participant filter which removes participants based upon their violations of the controlled conditions of the experiment. This step is mostly used to remove participants who were presented with an incomplete recommendation at some point during the experiment. Seeing a shorter recommendation than on previous pages might influence the participant's behaviour. As such, participants who encountered this situation are excluded.

F1. First Experiment

The second and third filtering steps are applied separately to the results of each experiment. Below are the filtering results for the first experiment.

IP Filter

Table 33 below provides an overview of all non-human visitors which were filtered from the experiment at the IP-address level, based upon their behaviour.

IP Address	Hostname	Recommendations	Participants
188.165.136.18	188-165-136-18.ovh.net	523	523
194.109.159.57	ia200128.eu.archive.org	513	513
194.176.105.54	inetgw-69-sec.nhs.uk	499	435
198.185.24.201	d2-prod-gw.lexisnexis.com	10.874	1.177
218.28.77.7	pc0.zz.ha.cn	347	220
64.38.3.50	cache1.linkpimp.net	315	317
65.98.224.5	cust-65-98-224-5.static.o1.com	302	303
72.14.193.66	72.14.193.66	747	728
74.125.16.66	74.125.16.66	1.526	1.475
74.125.74.193	74.125.74.193	427	415
78.137.163.133	ip-78-137-163-133.dedi.digiweb.ie	212	220
87.238.84.64	87-238-84-64.amazon.com	1.282	1.301
89.149.244.21	89.149.244.21	356	270
115.49.34.169	hn.kd.ny.adsl	2.715	397
115.49.91.171	hn.kd.ny.adsl	216	109
115.59.74.206	hn.kd.ny.adsl	308	247

Table 33: List of non-human visitors identified by behavioural filtering at IP level

Further results of the IP filter are the exclusion of 5.599 visitors based on them not accepting the experimentation cookie. This entails around 6% of the total number of participants is excluded for not accepting the experimentation cookie.

Participant Filter

The participant filter excluded 4.093 participants who during their visit saw one or more incomplete recommendations.

Figure 44 and *Figure 45* on the next page provide an overview of the effects of applying the second and third filtering steps on the dataset of the first experiment. These figures add additional detail to the overviews provided in *Figure 12* and *Figure 14* of *Chapter 4.4* respectively.

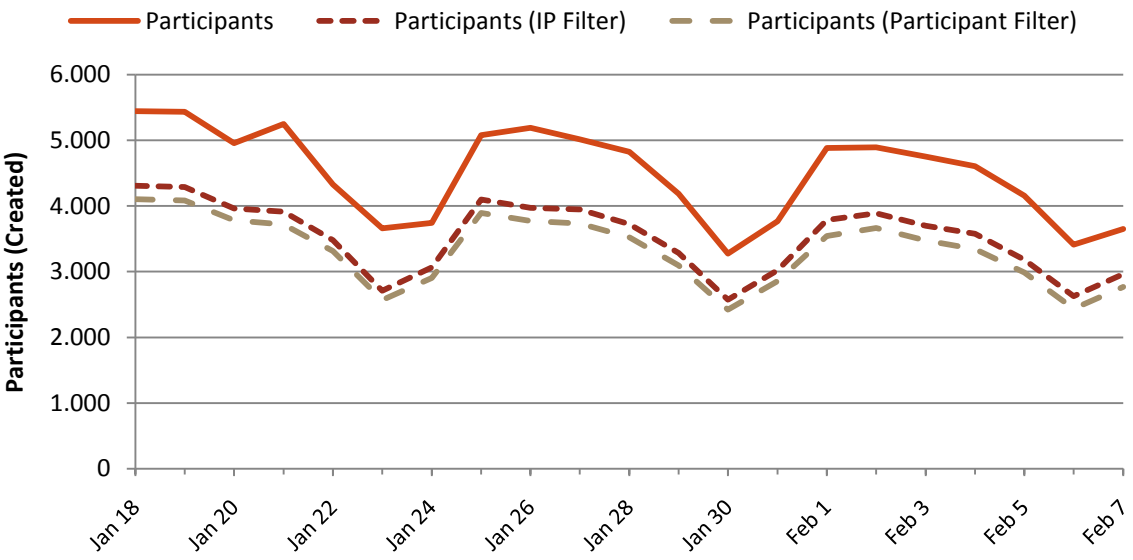


Figure 44: Daily number of participants created during the experimentation interval, before and after filtering

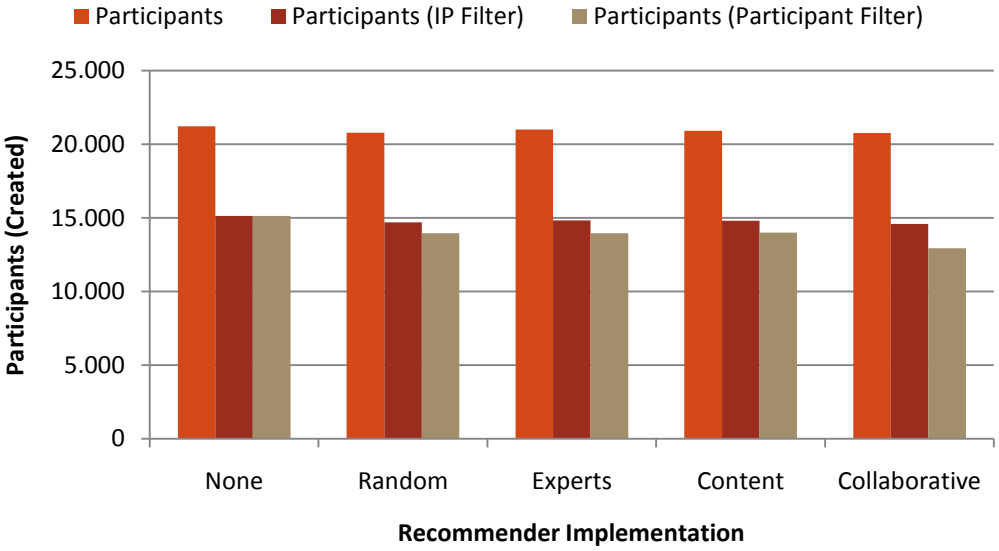


Figure 45: Total number of participants for each of the experiment groups

F2. Contextual Factors Experiment

The second and third filtering steps are applied separately to the results of each experiment. Below are the filtering results for the second experiment.

IP Filter

Table 34 below provides an overview of all non-human visitors which were filtered from the experiment at the IP-address level, based upon their behaviour.

IP Address	Hostname	Recommendations	Participants
194.176.105.54	inetgw-69-sec.nhs.uk	302	242
198.185.24.201	d2-prod-gw.lexisnexis.com	1.891	20.156
38.99.98.4	h-98-4.scoutjet.com	514	478
38.111.147.86	38.111.147.86	445	455
72.14.193.65	72.14.193.65	338	526
74.125.16.68	74.125.16.68	650	931
74.125.74.132	74.125.74.132	261	367
115.49.92.108	hn.kd.ny.adsl	396	170
115.49.94.110	hn.kd.ny.adsl	589	40

Table 34: List of non-human visitors identified by behavioural filtering at IP level

Further results of the IP filter are the exclusion of 3.247 visitors because they did not accept the experimentation cookie. This entails around 6% of the total number of participants were excluded. This fraction is equal to the fraction uncovered during the first experiment.

Participant Filter

The participant filter excluded 7.474 participants who saw one or more incomplete recommendations during their visit.

Figure 46 and *Figure 47* on the next page provide an overview of the effects of applying the second and third filtering steps on the dataset of the second experiment. These figures add additional detail to the overviews provided in *Figure 23* and *Figure 25* in *Chapter 6.4* respectively.

The large difference between the number of participants before and after the IP filter, in comparison with the first experiment, is explained by looking at *Table 34*. There is a single address, *d2-prod-gw.lexisnexis.com*, responsible for over 20.000 participants during the second experiment. This number alone is more than the total number of participants filtered during the first experiment. It explains the large drop seen between the “Participants” and “Participants (IP Filter)” datasets in both figures on the next page.

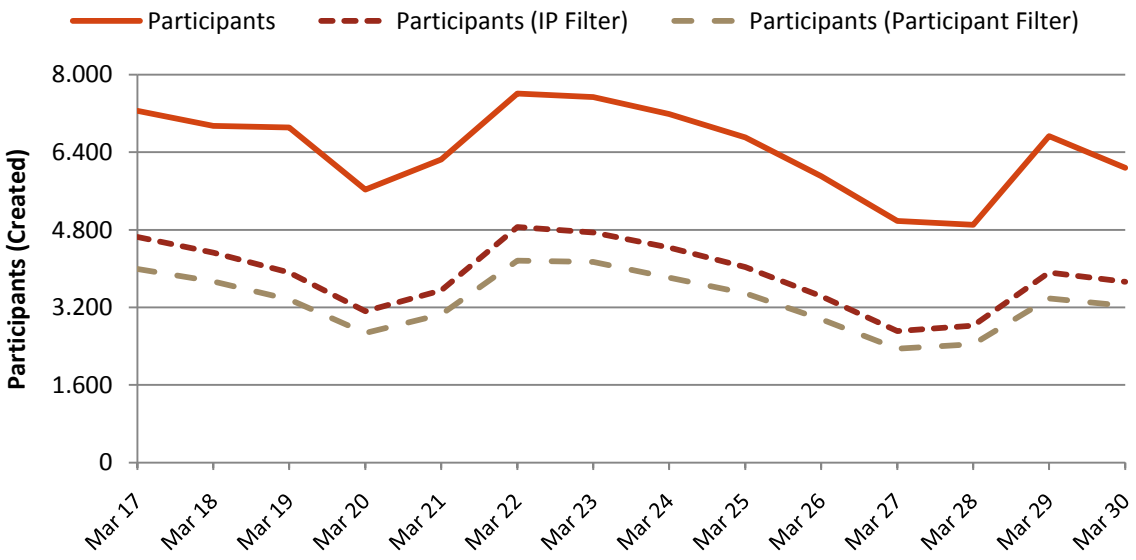


Figure 46: Daily number of participants created during the second experiment, before and after filtering

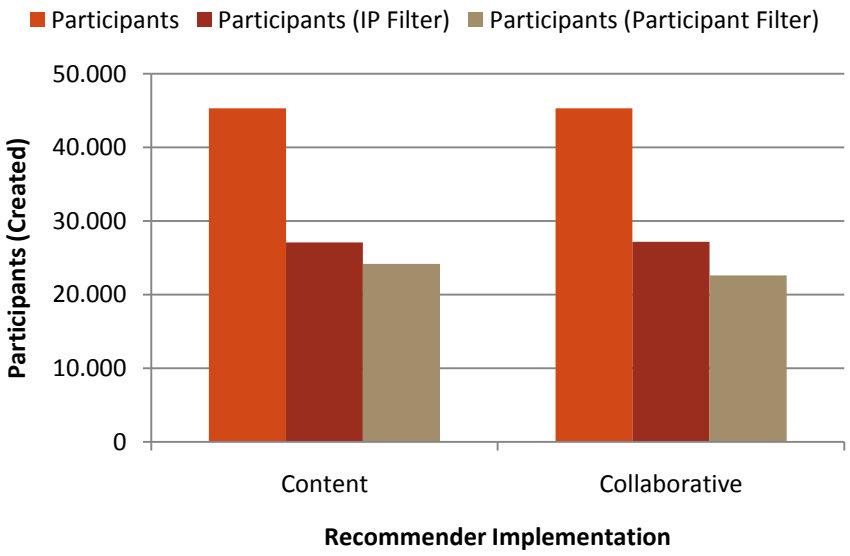


Figure 47: Total number of participants for each of the experiment groups during the second experiment

G. Validation

This appendix provides detailed results for the validation of the first experiment, as described in *Chapter 4.5*, and the second experiment, as described in the *Chapter 6.5*. A further discussion concerning the validation procedure is available in *Chapter 2.5.2*.

G1. First Experiment

For the first experiment, two validation steps are performed. Firstly, the results are compared against results generated by a third-party statistics package, Webalizer. Secondly, a purpose-build web server log-file harvester is used to validate the results of the experiment.

Comparison with Webalizer

For the initial validation step the visitor statistics for the MastersPortal.eu website generated by its third-party statistics package, Webalizer, are compared against the results of the experiment in an attempt to verify both sets of data show a similar visitor trend.
Looking at the number of visits and IP addresses as reported by Webalizer in *Figure 48* we see the familiar weekly pattern emerge again. Furthermore, we again see a strong correlation between the number of visits and the number of IP addresses recorded.

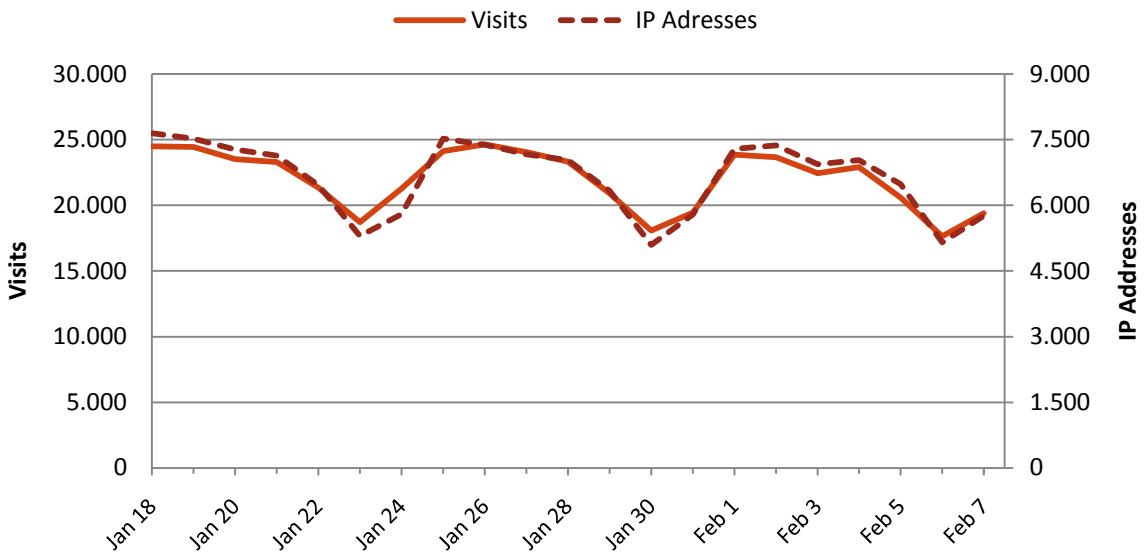


Figure 48: Webalizer statistics for the first experiment

In the following graph, *Figure 49*, the results for the first experiment, expressed by the number of sessions started and the number of participants created, are compared against the number of IP addresses as recorded by Webalizer.

We see a strong correlation between the number of IP addresses reported by Webalizer and the statistics extracted from the first experiment. No strange peeks or diversions, pointing towards potential issues with the experiment, are present in the graph. Concluding we can state the comparison of the statistics generated by Webalizer against the results of the experiment indicates the experiment results are valid.

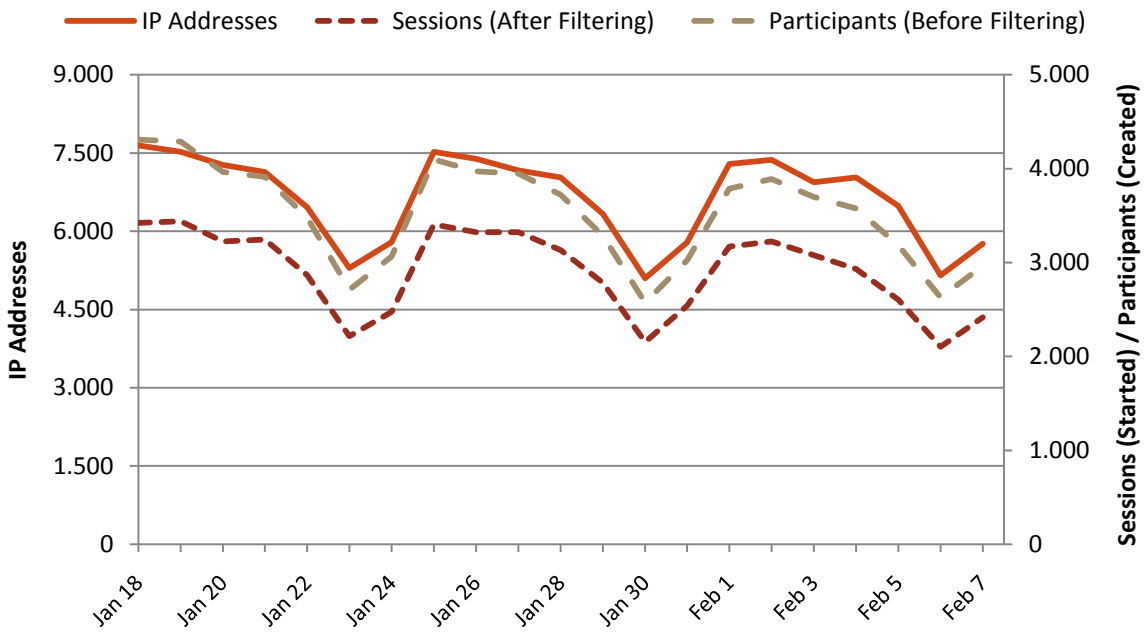


Figure 49: Webalizer statistics compared to the first experiment results

Apart from comparing the Webalizer figures against the figures for the first experiment, the diagram in *Figure 49* also provides verification of the assumption that the filtering of participants has no effect on the overall visitor trend and chronological distribution of participants.
The number of sessions listed in the graph is the number of sessions *after* all filtering steps are applied. The number of participants listed is the number of participants *before* any of the filtering is applied. Since both lines follow a very similar trend it is safe to assume that filtering did not introduced any inconsistencies.

Comparison with Thesis Harvester

For the second part of the validation the custom log-file harvester, constructed to analyse the visitor statistics in the introduction of this thesis, was repurposed. This newly repurposed harvester scanned the web server log-files and processed them with the purpose of gathering the same metrics as measured by the experimentation code.
While the previous validation step shows that the results of the experiment are globally in line with the overall visitor trend, this analysis attempts to assert the results presented are in line with what is recorded by the web server itself.

It is important to note that the harvested results are an approximation of what actually happened. Due to the stateless nature of the HTTP protocol, not all information required can be retrieved from the web server log-files. For example, the harvester basis its distinction of sessions on difference in IP addresses encountered; the experiment identifies participants using cookies, tracking them through the experiment.

In *Figure 50* the number of sessions started both directly taken from the experiment and gathered by the harvester are compared. Both datasets show a strong correlation.

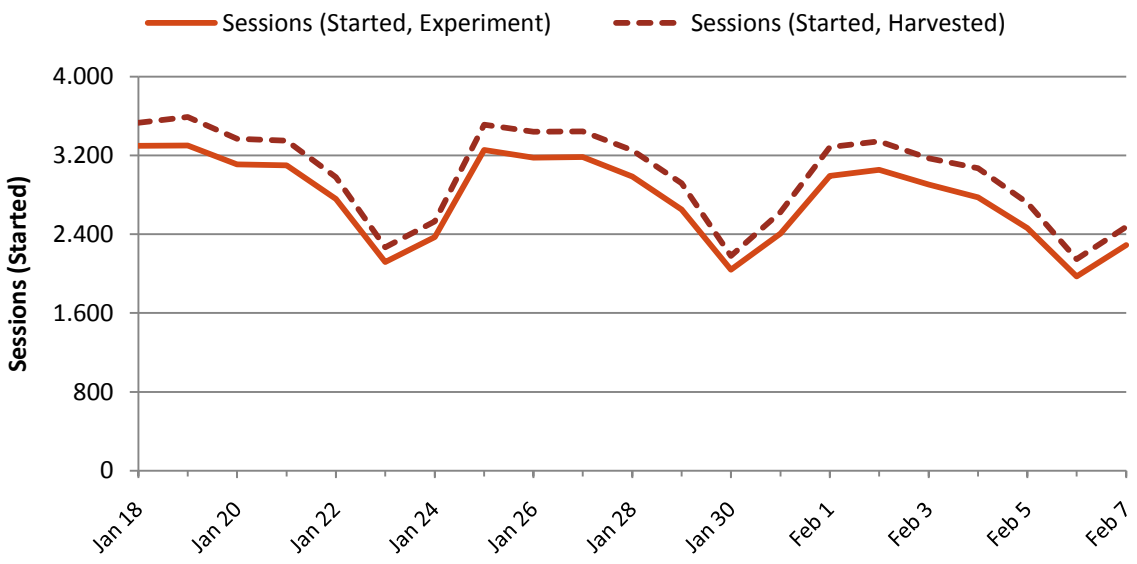


Figure 50: Experiment statistics compared against harvested results.

The number of sessions started as reported by the experiment is slightly lower than the number reported by the harvester. This is caused by the fact that the final filtering steps applied during the experiment are applied at the participant level. Since the harvester has no knowledge of participants, it could not apply these filtering steps and thus did not exclude some of the participants which were excluded during the experiment.

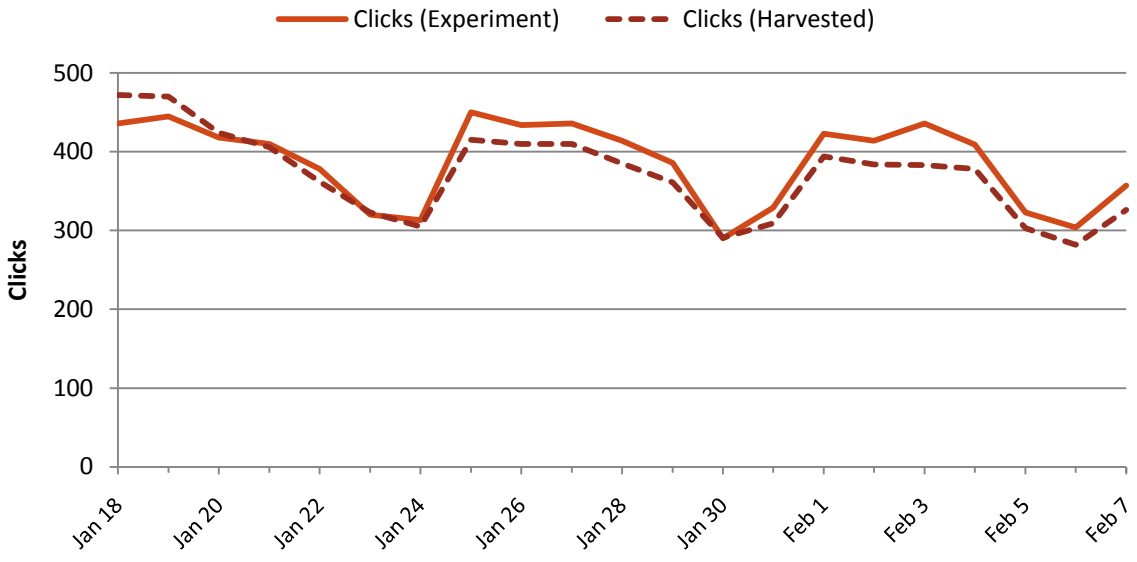


Figure 51: Experiment clicks compared against harvested clicks.

In *Figure 51* the number of clicks as reported by the experiment is compared to the number of clicks reported by the harvester. As the harvester cannot actually detect clicks, its conclusions are based on inferring "clicks" from the page requests found in the web server log-files. The number of clicks as reported by the harvester is thus an approximation. From the graph it is clear that both datasets again strongly correlate.

G2. Second Experiment

For the second experiment, one validation step is performed. The results of the second experiment are compared against results generated by the third-party Webalizer statistics software.

Comparison with Webalizer

The visitor statistics for the MastersPortal.eu website generated by its third-party statistics package, Webalizer, are compared against the results of the experiment in attempt to verify both sets of data point towards a similar visitor trend.

Looking at the number of visits and IP addresses as reported by Webalizer in *Figure 52*Figure 48 we see the familiar weekly pattern emerging again. Furthermore, we see a strong correlation between the number of visits and the number of IP addresses recorded.

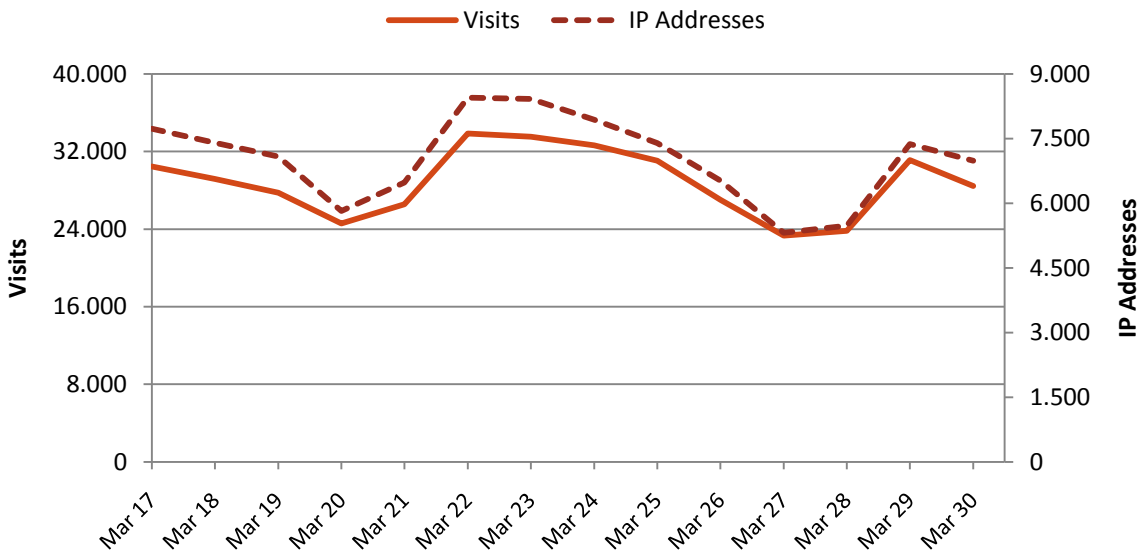


Figure 52: Webalizer statistics for the second experiment

In *Figure 53* on the next page the results for the first experiment, expressed by the number of sessions started and the number of participants created, are compared against the number of IP addresses as recorded by Webalizer.

Just as with the first experiment we see a strong correlation between the number of IP addresses reported by Webalizer and the statistics extracted from the experiment. No strange peeks or diversions, pointing towards potential issues with the experiment, are present in the graph. Concluding we can state the comparison of the statistics generated by Webalizer against the results of the experiment indicates the experiment results are valid.

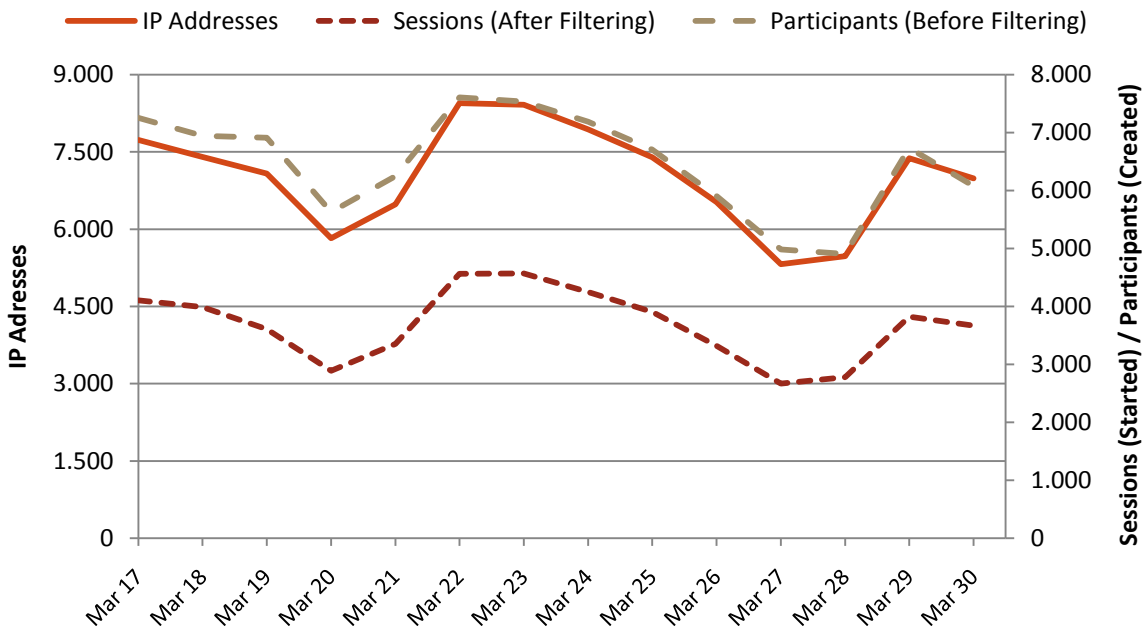


Figure 53: Webalizer statistics compared to the second experiment results

The number of sessions started, as listed in *Figure 53*, is the number of sessions started *after* all filtering is applied. The number of participants listed is the number of participants *before* any of the filtering steps are applied. This provides an indication the filtering applied has not introduced any anomalies in the data.

In Figure *53* we see a rather large difference between the number of participants created, before filtering, and the number of sessions started, after filtering. An explanation for this fact is provided in *Appendix F2*.

H. Replay Procedure Implementation Details

This appendix shortly describes the replay procedure used to test the feasibility of the adaptive recommender system. The goal of this appendix is to provide an overall insight into the procedure used. This allows a similar procedure to be constructed for future experiments.

H1. Replay of the Experiment

The replay of the experiment's data consists of two iterations. Before starting the first iteration, the results of the experiment are filtered using a procedure similar to that applied prior to the main analyses of the experiments' results as discussed in this thesis. Again, only sessions starting with a referral from Google are taken into account.

During the first iteration of the replay each recommendation generated over the course of the experiment is passed into the replay system. Recommendations are sorted in ascending order based on their timestamp. For each session encountered it is determined if the sessions bounced. After the first iteration the replay timer is reset to zero.
During the second iteration the adaptive system makes its decisions. At the start of each session the scoring system is invoked. It computes a score for each individual context and combines these scores into a single overall score for the session. Details on the scoring procedure are provided in *Chapter 7.2*. The scoring decisions for each session are saved to an external file. Further analyses of the adaptive system are executed using Microsoft Excel.

The following four contexts are used during the replay.

1. Geographical Origin
2. Google Query
3. Academic Discipline
4. Screen Resolution

Further details on each of these contexts are provided in *Chapter 5.2*. The "Google Query" context is discussed in further detail in the next section of this appendix. This context is less straightforward than the other three contexts and as such required a purpose-build analysis which is discussed in the next section.

H2. Contextual Factor "Google Query"

Classifications based upon the Google query context are more complicated than those for the other three contexts. It is not a matter of reading a simple property from either the visitor or the programme viewed; the query-string entered by the visitor on Google needs to be interpreted. A simple textual analysis procedure was thus created to classify visitors based upon the context. This appendix details the procedure.

Both the query-string provided by the visitor and all strings used for comparison are filtered of stop words and have Porter's stemming algorithm (Porter, 2006) applied to their terms. All strings are furthermore transliterated into ASCII to remove diacritics and converted to lower-case. Finally, the order of terms in all strings is considered irrelevant.

The first part of the classification procedure is to identify if the query-string point towards a general interest in an academic education. For this purpose a set of triggers is devised. The selection of these triggers is based upon a qualitative analysis of the search-strings entered by previous visitors.
If one or more of the triggers presented in *Table 35* is present in the search-string it is assumed the search-string points towards an "interest in a Master's degree".

"Interest in Master's Degree" Triggers

master	erasmus mundus	postgrad	program
masters	degree	graddip	programs
msc	post graduate	pgdip	programme
mba	grad dip	diploma	programmes
llm	post grad	school	

Table 35: Terms triggering the "Interest in Master's Degree" classification

Subsequently, a further classification of the participants with an "interest in a Master's degree" into one or more of the four sub classifications listed below is attempted.

1. Europe Interest
2. Country Interest
3. University Interest
4. Title-Match

Europe Interest

The "Europe interest" classification is made by scanning the search-string for the trigger "Europe". If it is present in the search-string it is assumed an interest in "studying in Europe" exists for the visitor.

Country Interest

The "country interest" classification is made by scanning the search-string for any of the country names present in the MastersPortal database. The list of countries names acting as triggers for this classification is provided in *Table 36*.
Two country names were manually added to the list: UK and Holland. A qualitative analysis of the search-strings shows that these two terms are used often instead of the proper names of the respective countries: the United Kingdom and The Netherlands.

"Country Interest" Triggers				
albania	estonia	italy	netherlands	spain
austria	finland	latvia	norway	sweden
belgium	france	lithuania	poland	switzerland
bosnia	germany	luxembourg	portugal	turkey
bulgaria	greece	macedonia	romania	ukraine
croatia	hungary	malta	russia	united kingdom
cyprus	iceland	moldova	serbia	uk
czech	ireland	monaco	slovakia	holland
denmark	isle man	montenegro	slovenia	

Table 36: Terms triggering the "Country Interest" classification

Visitors classified as having a "country interest" are assumed to be interested in studying in a European country for which the MastersPortal database provides Master's programmes.

University Interest

The "university interest" classification is made by scanning the search-string for the name of any of the over 1,500 universities present in the MastersPortal database. Due to its length, the full list is not included in this appendix.

This classification indicates the visitor has an interest in studying at a European university for which the MastersPortal database contains Master's programmes.

Title-Match

Finally, the title-match classification is based upon a match of the title of the Master's programme viewed by the visitor with the terms in his search-string. A match occurs if all terms in the search-string are present in the programme title, or vice-versa. Note that in this case discarding word order plays an important role in increasing the number of classification "hits".

The "title-match" classification indicates a visitor has a conceptual interest in the contents of the Master's programme he is referred to by Google.

I. Contextual Factors – Detailed Graphs

This appendix provides several additional graphs referred to in *Chapter 5*. The graphs in this appendix can be used as a visual aid to interpreting the combined graphs provided in the chapter.

I1. Geographical Origins

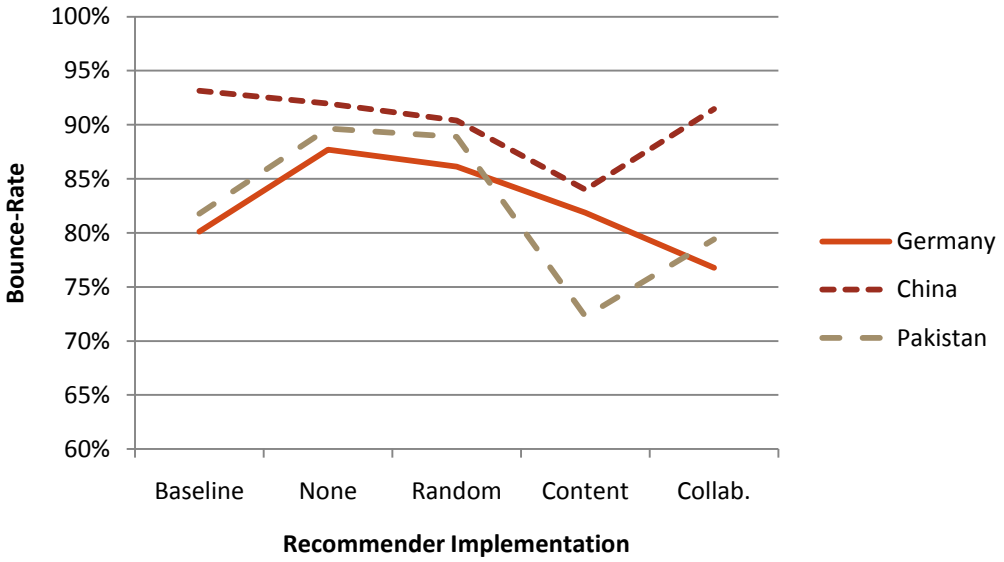


Figure 54: Bounce-rates for the first experiment when visitors are classified by country (1)

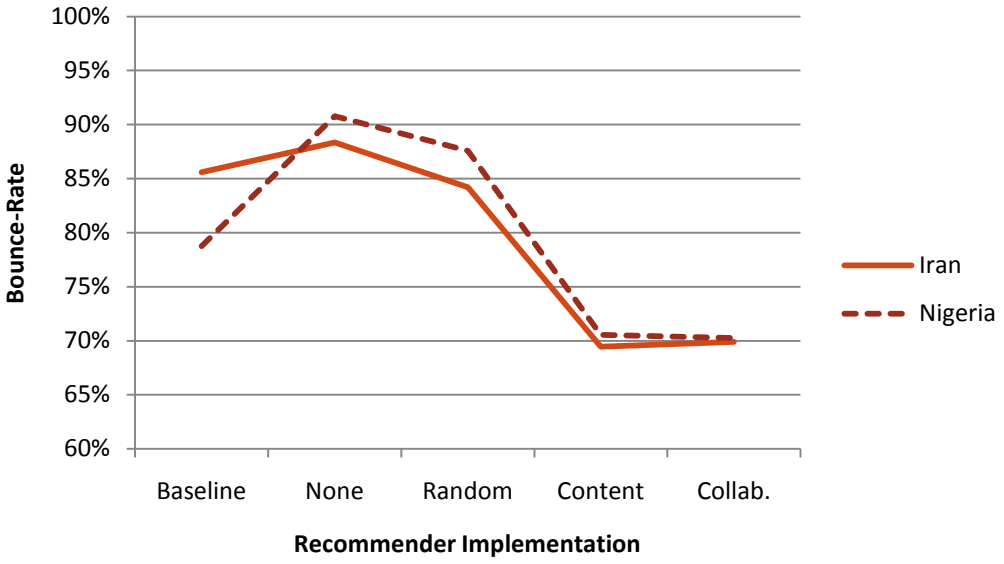


Figure 55: Bounce-rates for the first experiment when visitors are classified by country (2)

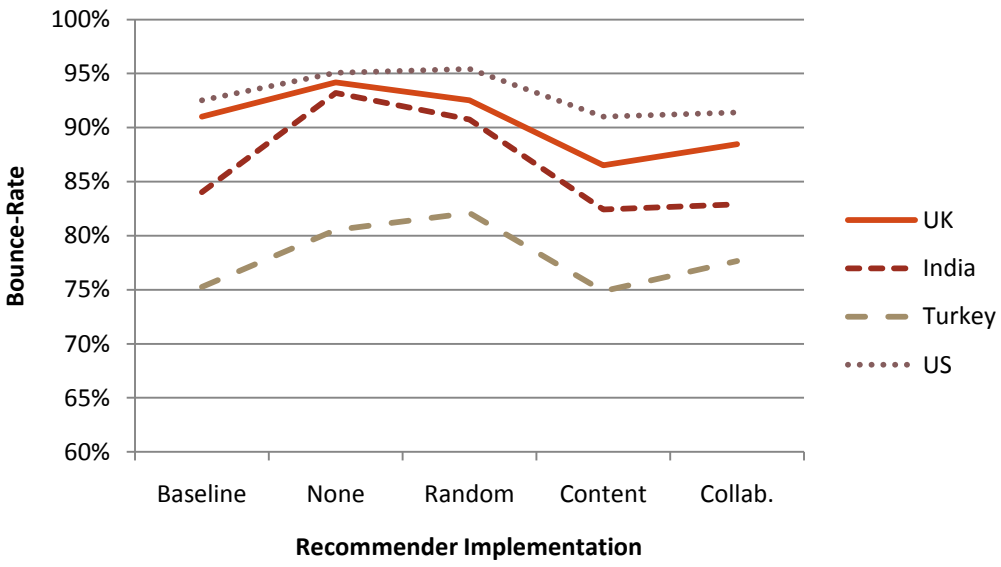


Figure 56: Bounce-rates for the first experiment when visitors are classified by country (3)

I2. Academic Disciplines

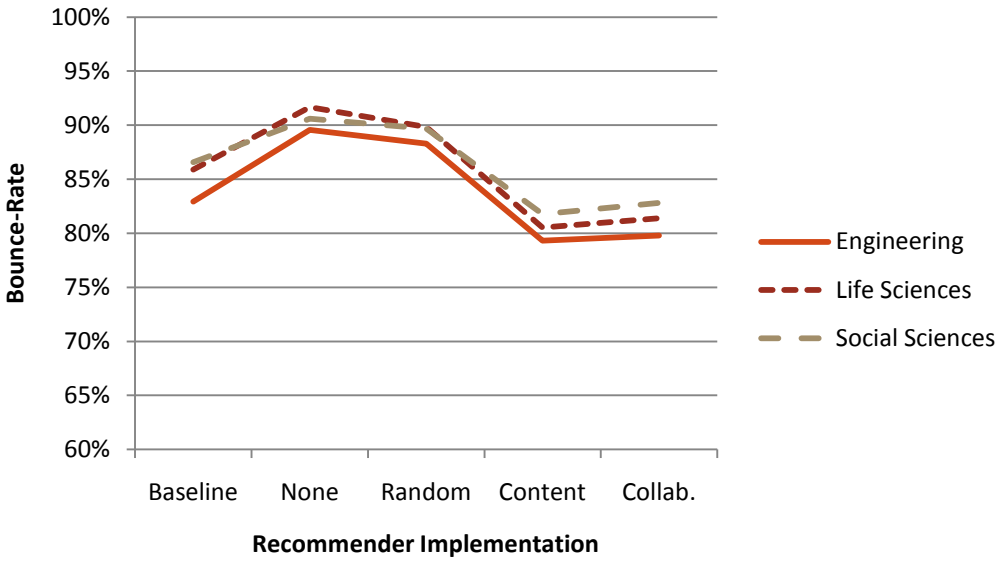


Figure 57: Bounce-rates for the first experiment when visitors are classified by academic discipline (1)

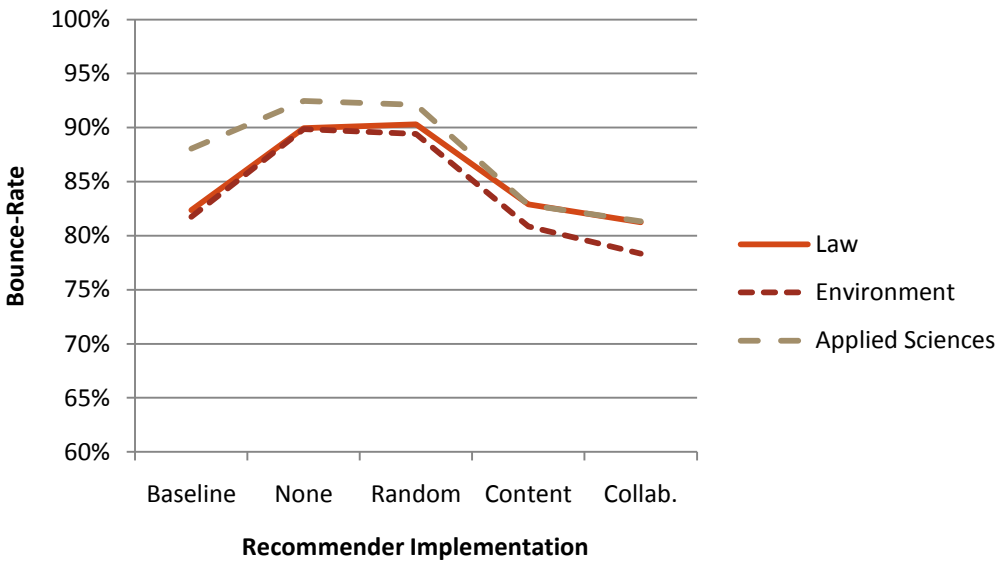


Figure 58: Bounce-rates for the first experiment when visitors are classified by academic discipline (2)

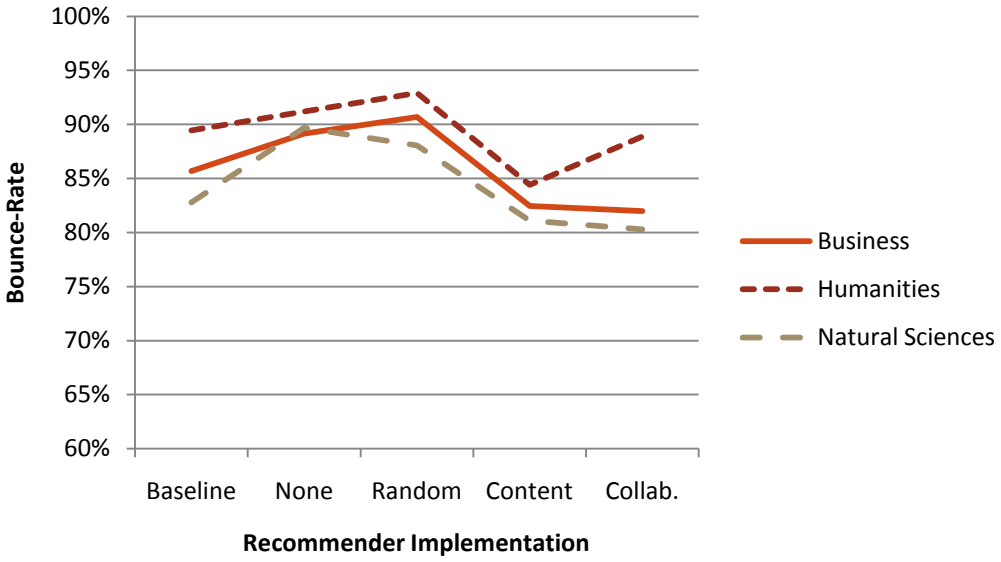


Figure 59: Bounce-rates for the first experiment when visitors are classified by academic discipline (3)

J. Feasibility of an Adaptive Recommender System – Detailed Graphs

The graphs in this appendix accompany *Chapter 7.4*. They provide detailed overviews of the effects within each contextual category.

J1. Geographical Origin

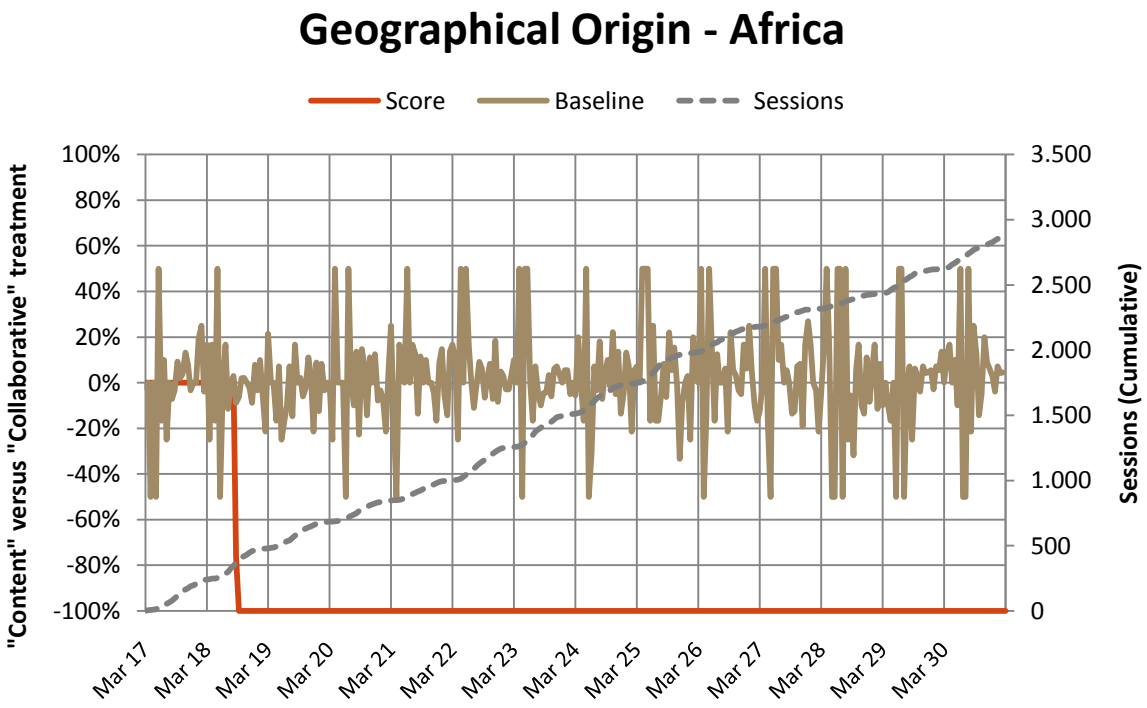


Figure 60: Geographical Origin – Africa

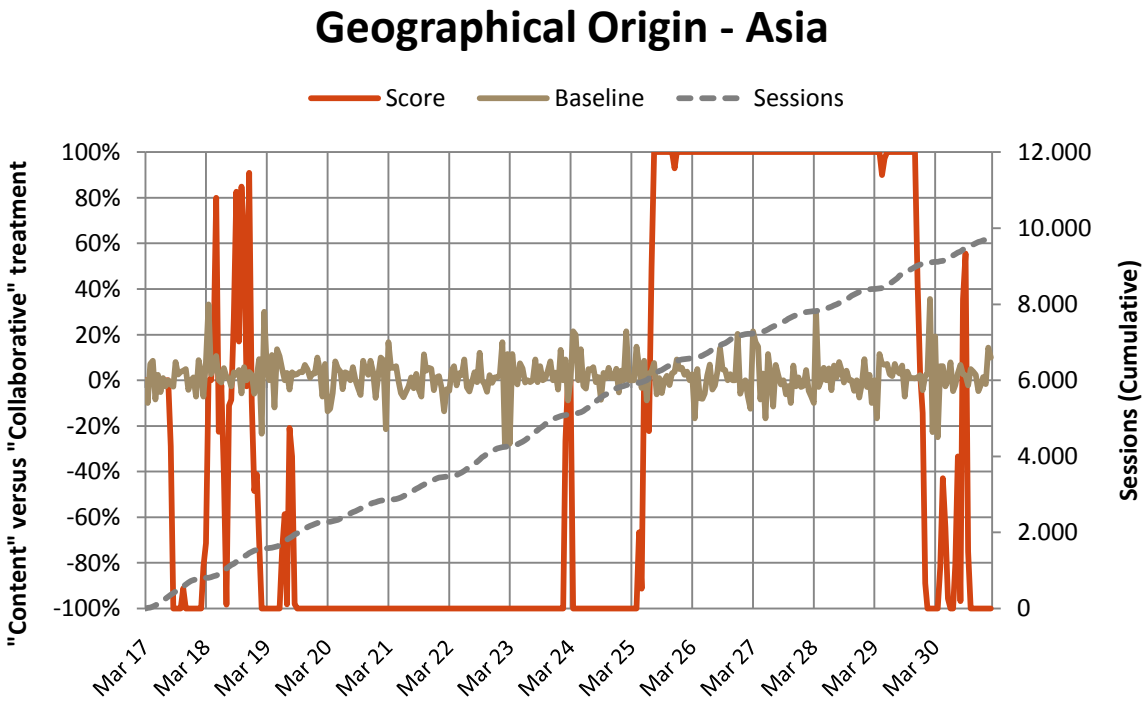


Figure 61: Geographical Origin – Asia

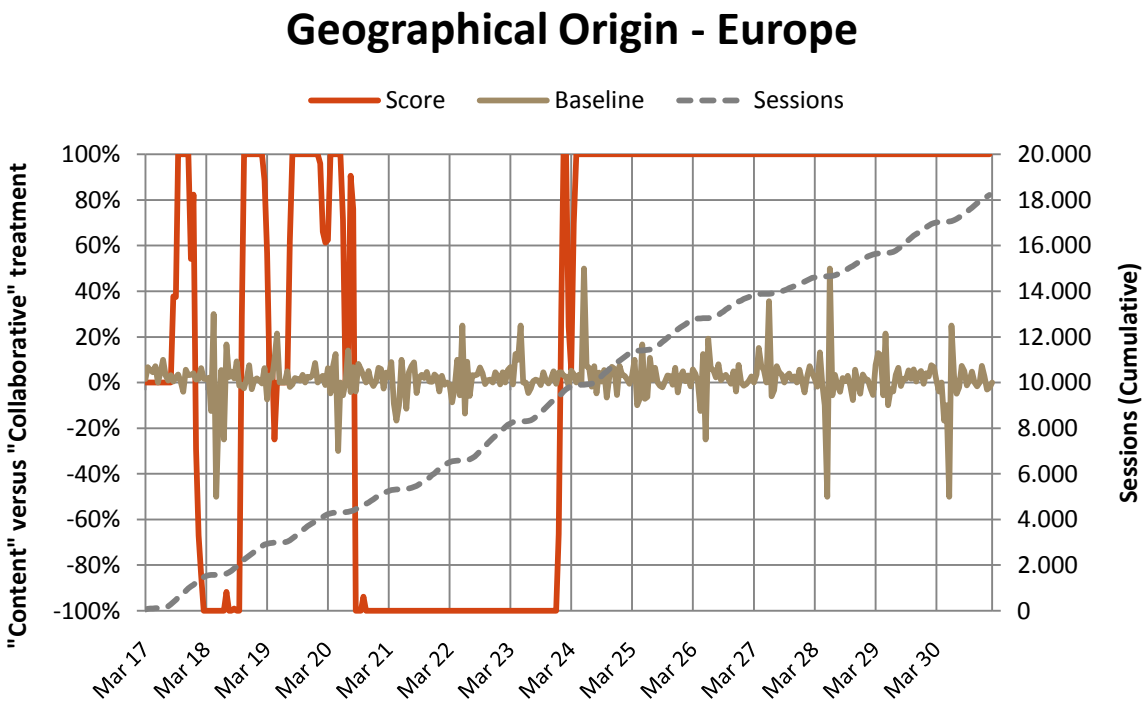


Figure 62: Geographical Origin – Europe

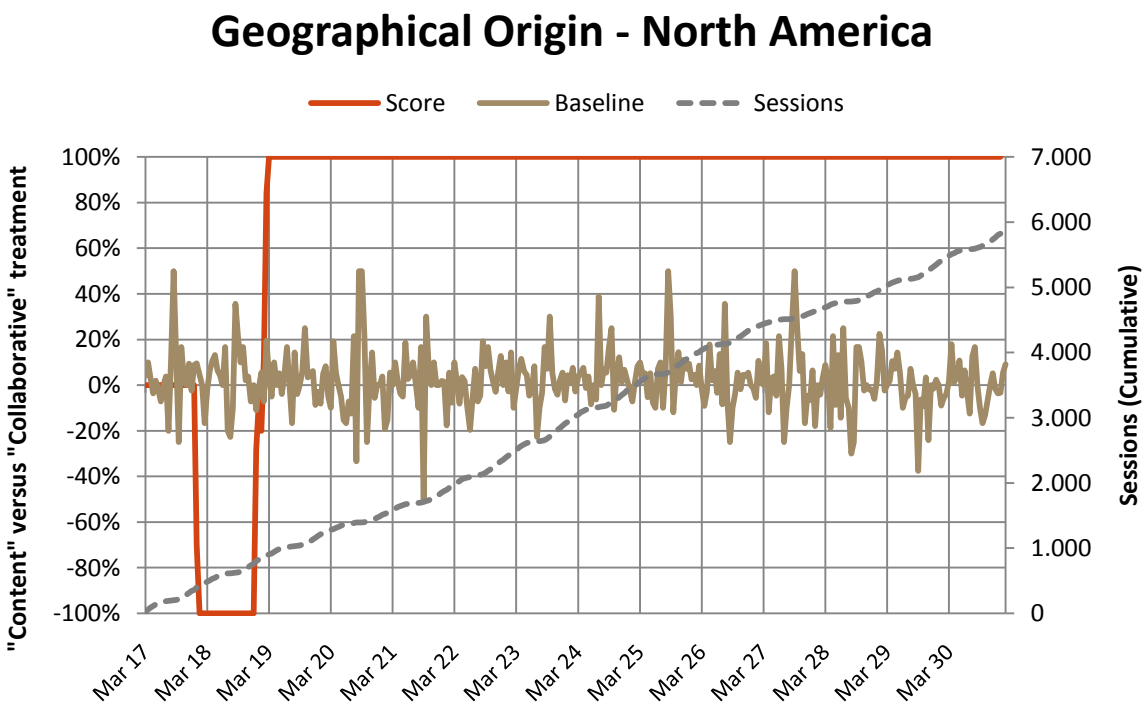


Figure 63: Geographical Origin – North America

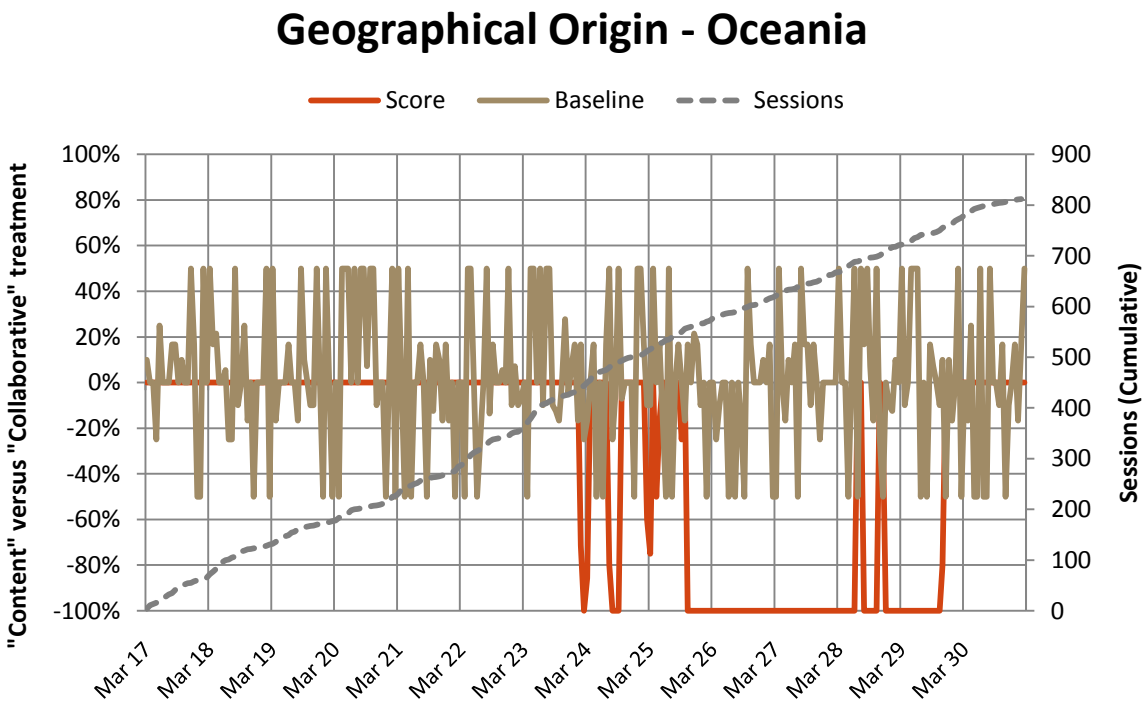


Figure 64: Geographical Origin – Oceania

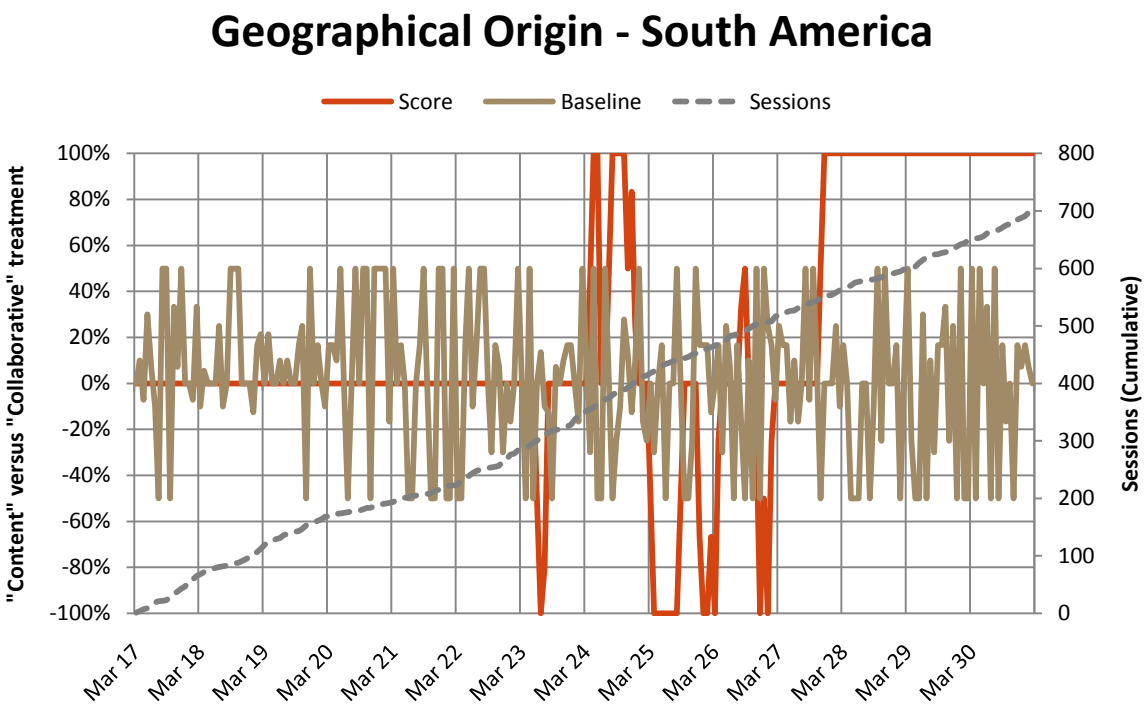


Figure 65: Geographical Origin – South America

J2. Google Query

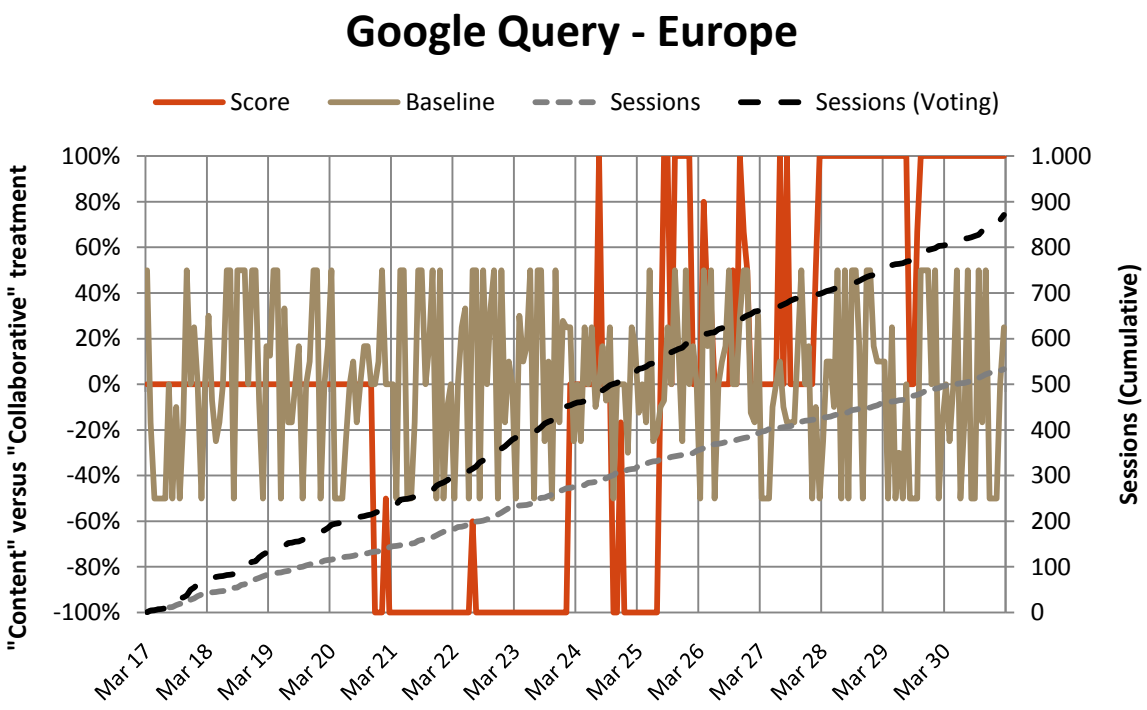


Figure 66: Google Query – Europe

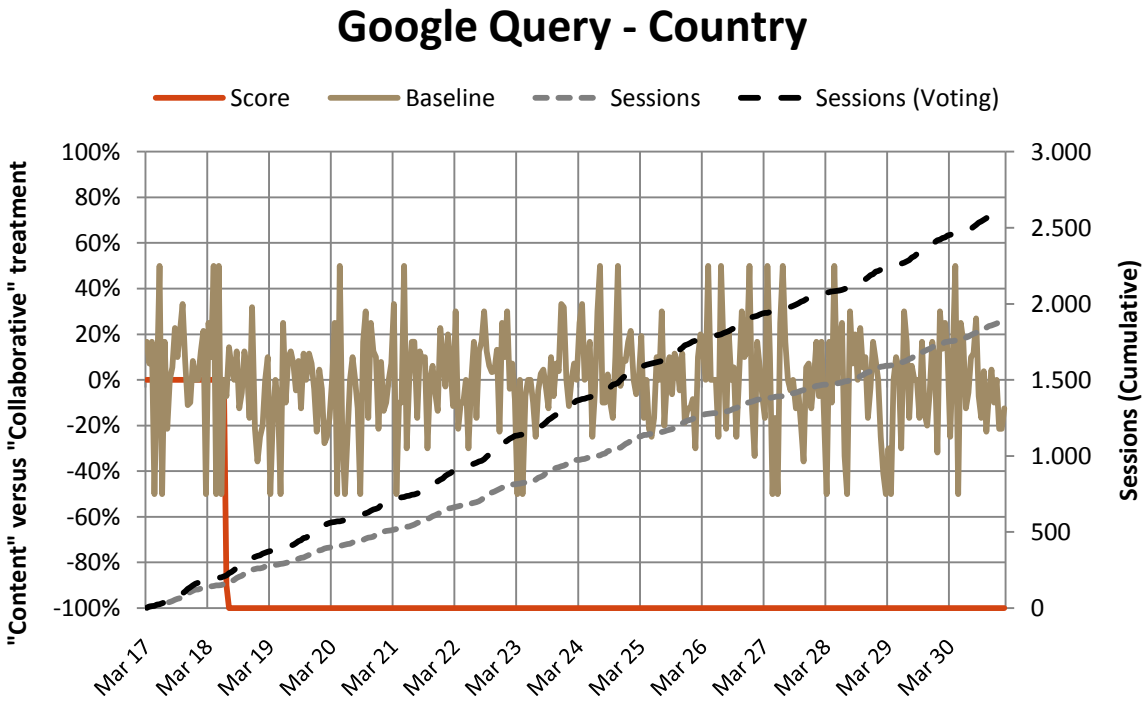


Figure 67: Google Query – Country

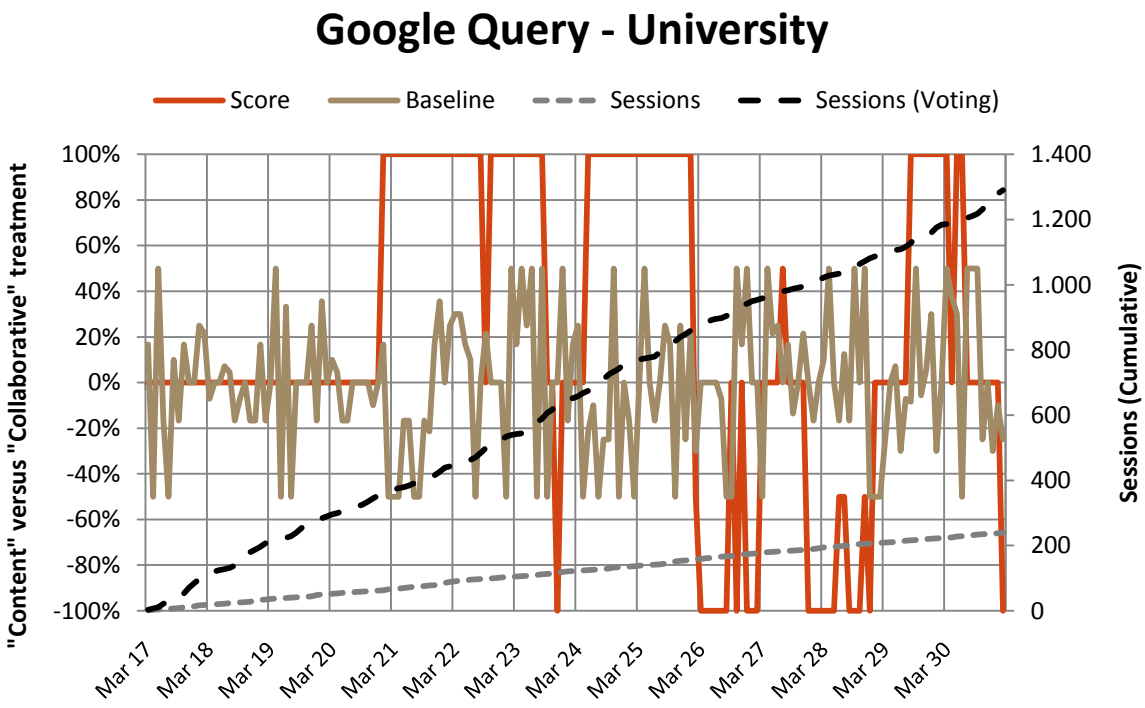


Figure 68: Google Query – University

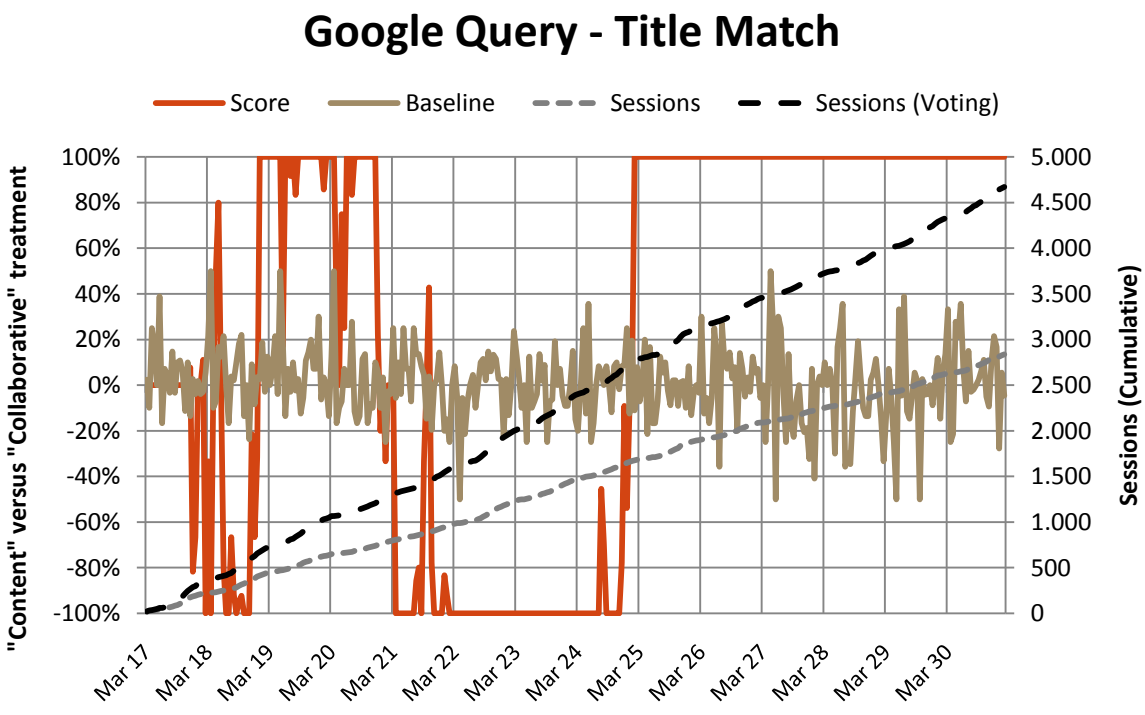


Figure 69: Google Query – Title Match

J3. Academic Discipline

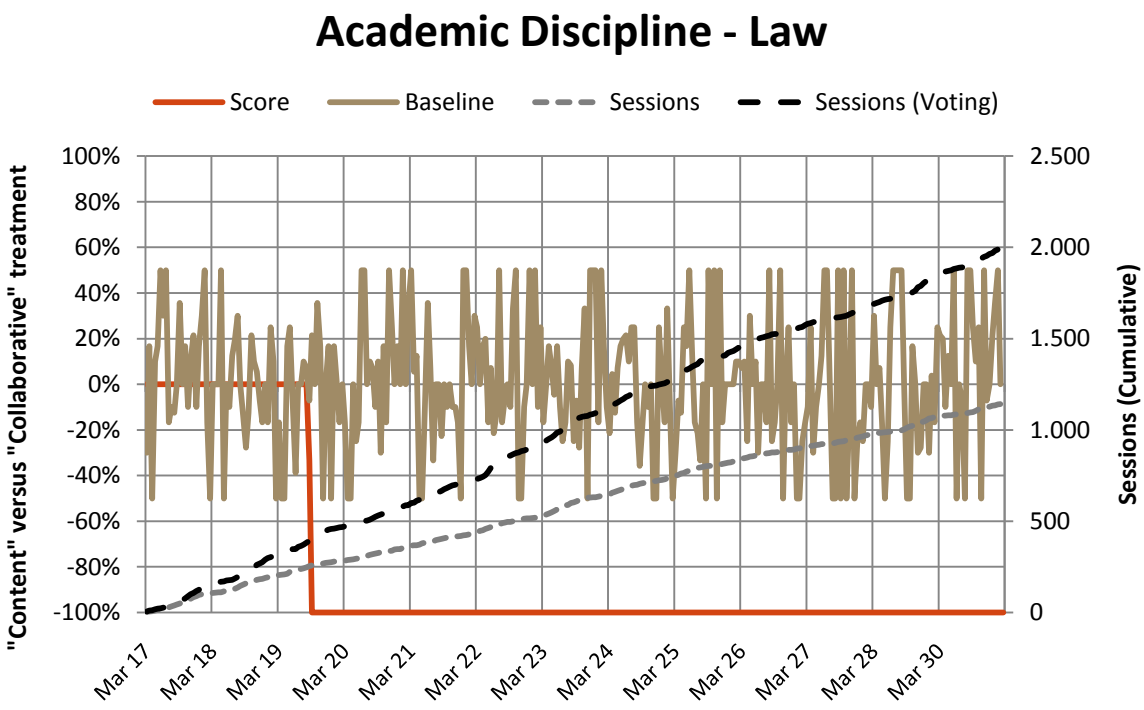


Figure 70: Academic Discipline – Law

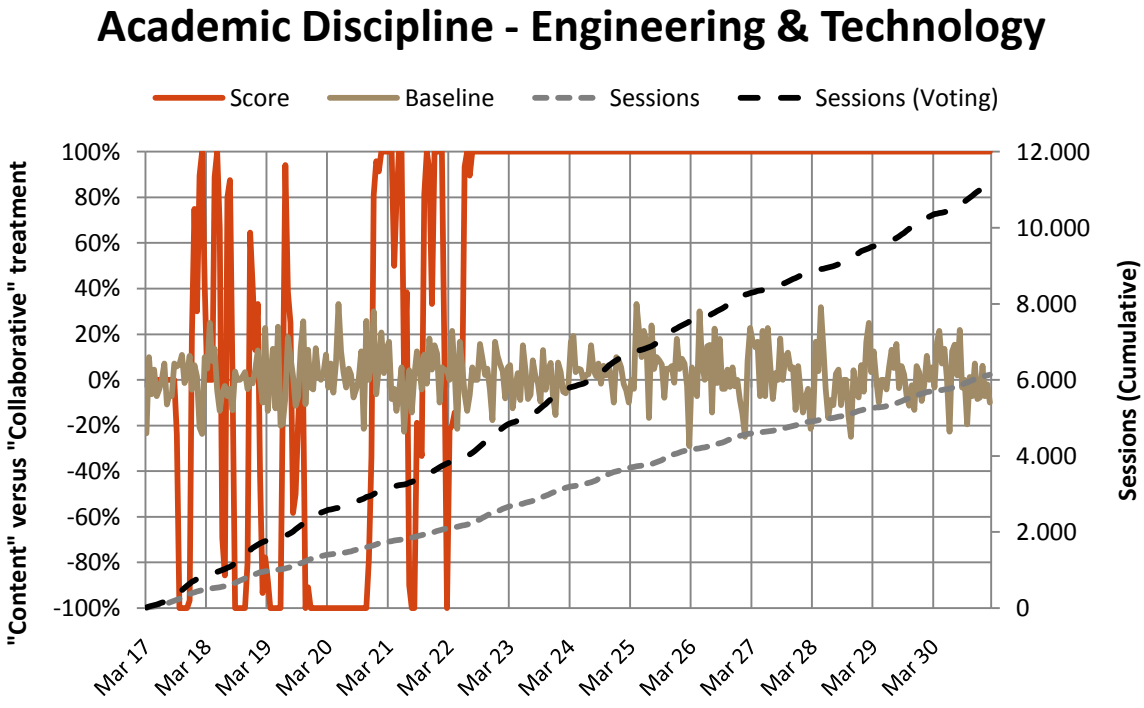


Figure 71: Academic Discipline – Engineering & Technology

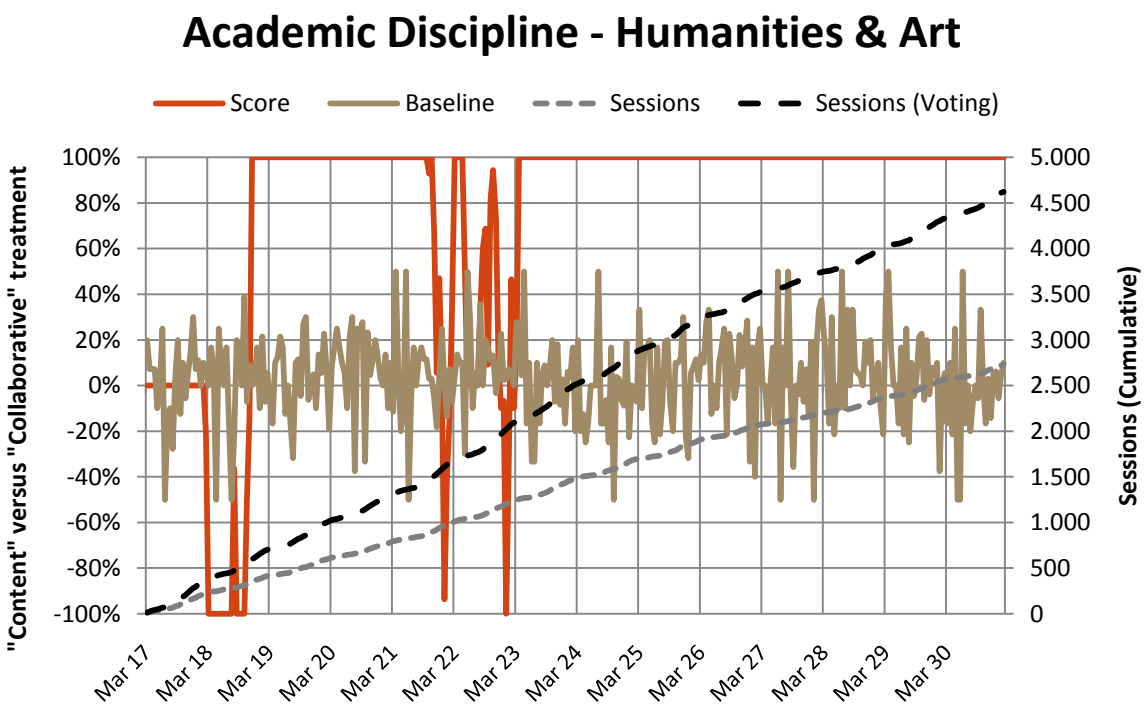


Figure 72: Academic Discipline – Humanities & Art

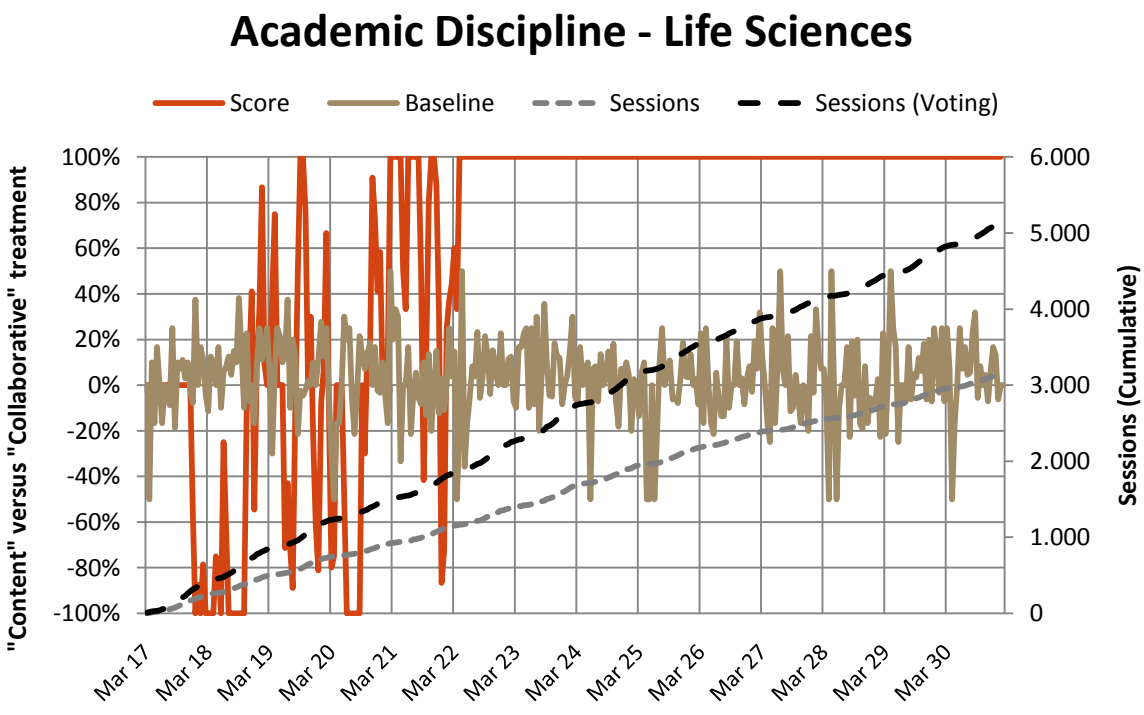


Figure 73: Academic Discipline – Life Sciences, Medicine & Health

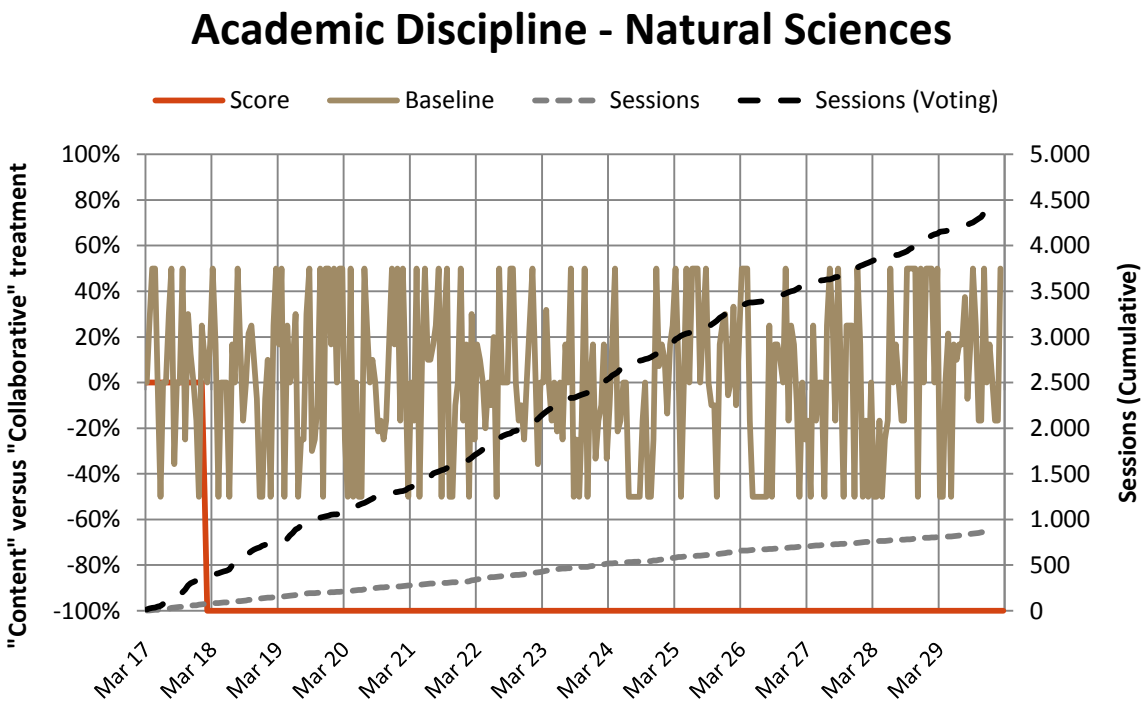


Figure 74: Academic Discipline – Natural Sciences

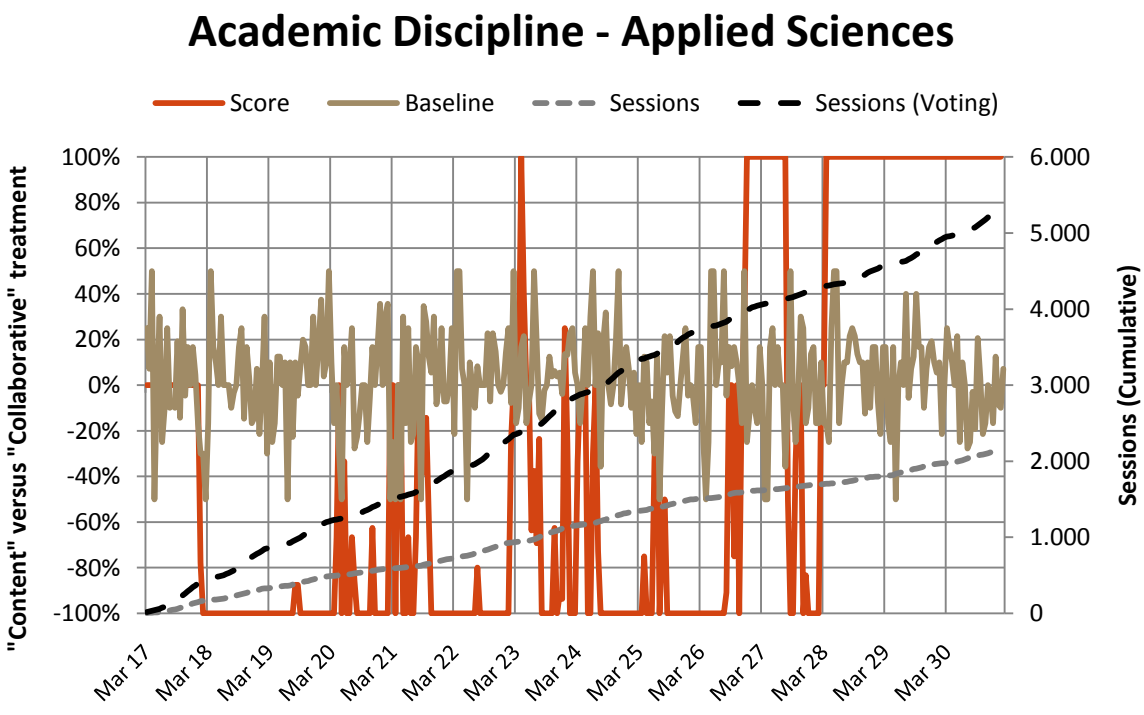


Figure 75: Academic Discipline – Applied Sciences, Professions & Arts

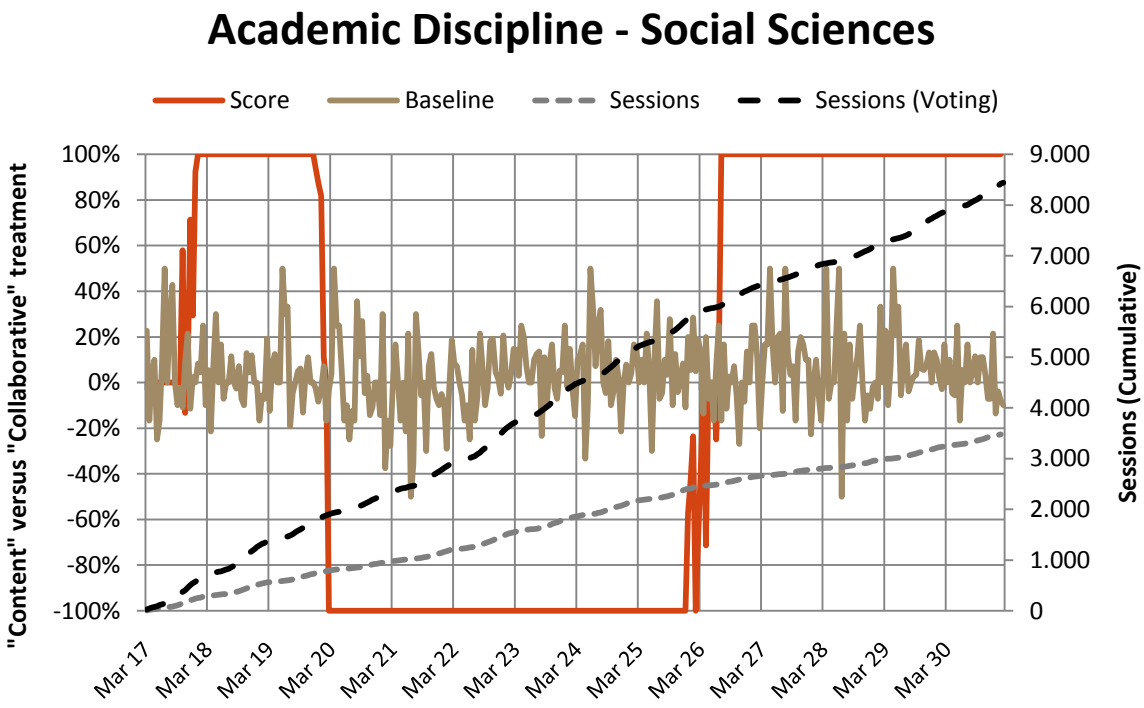


Figure 76: Academic Discipline – Social Sciences

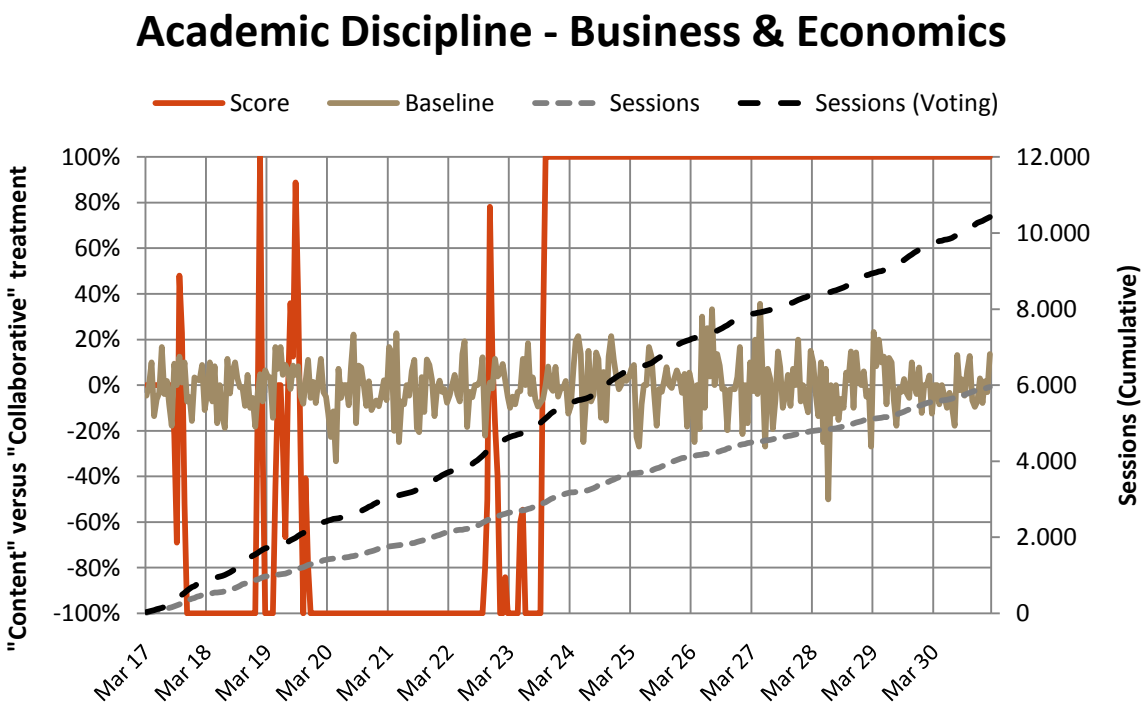


Figure 77: Academic Discipline – Business & Economics

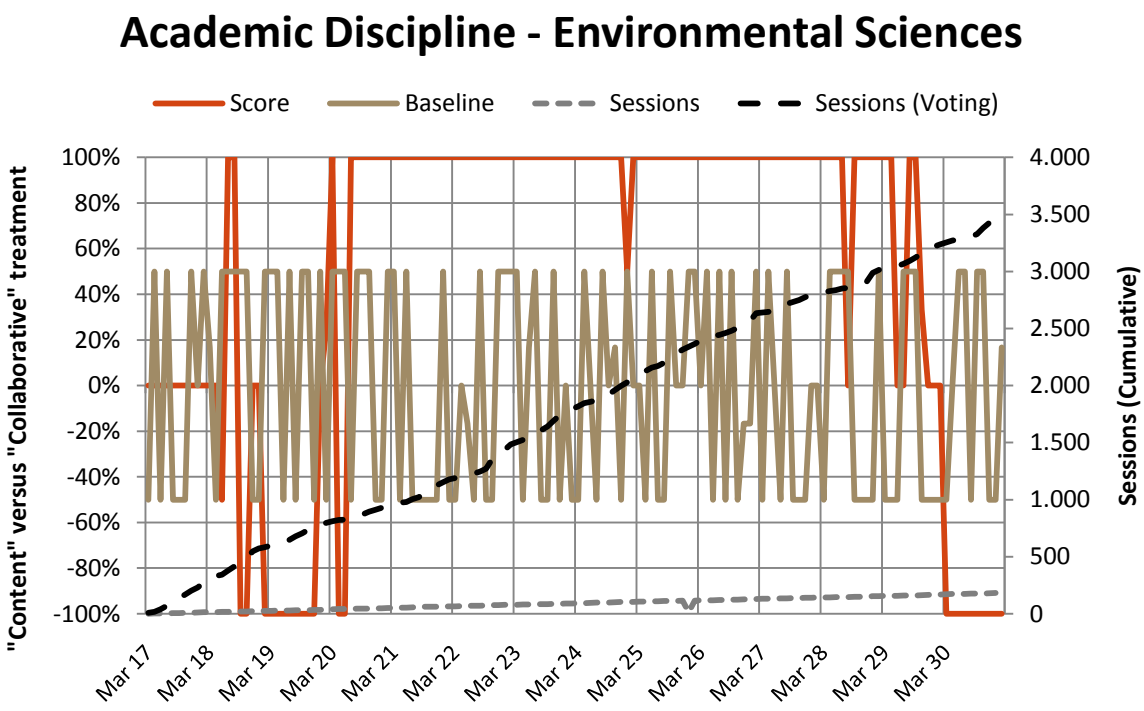


Figure 78: Academic Discipline – Environmental Sciences

J4. Screen Resolution

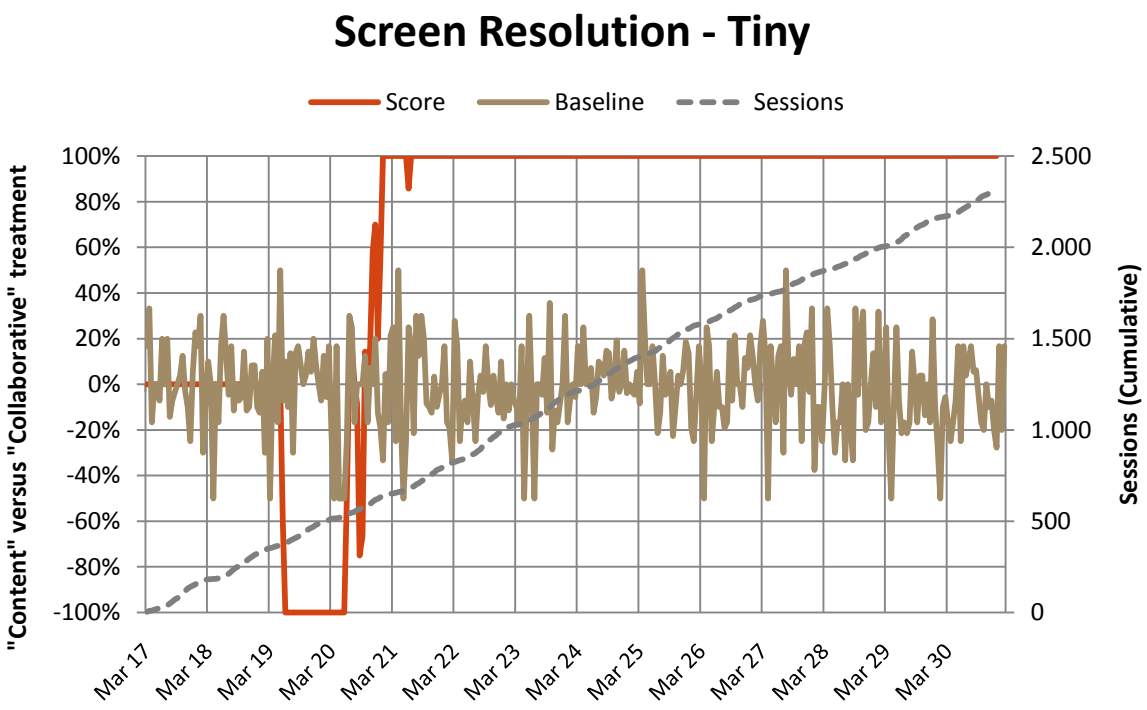


Figure 79: Screen Resolution – Tiny

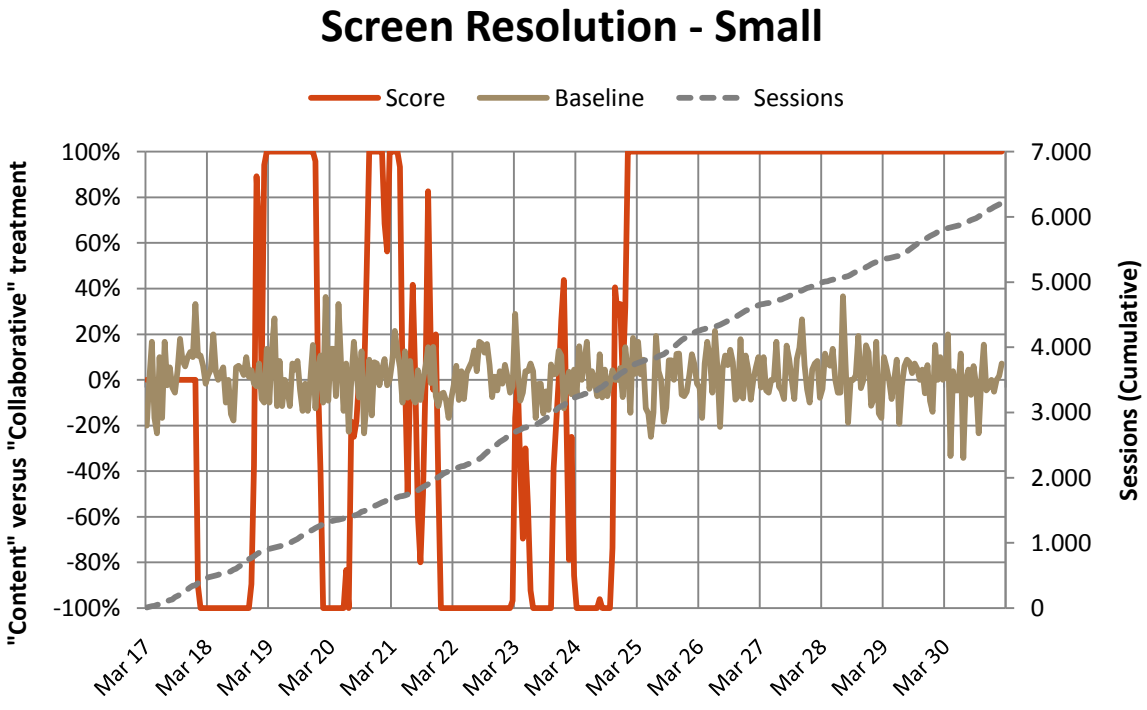


Figure 80: Screen Resolution – Small

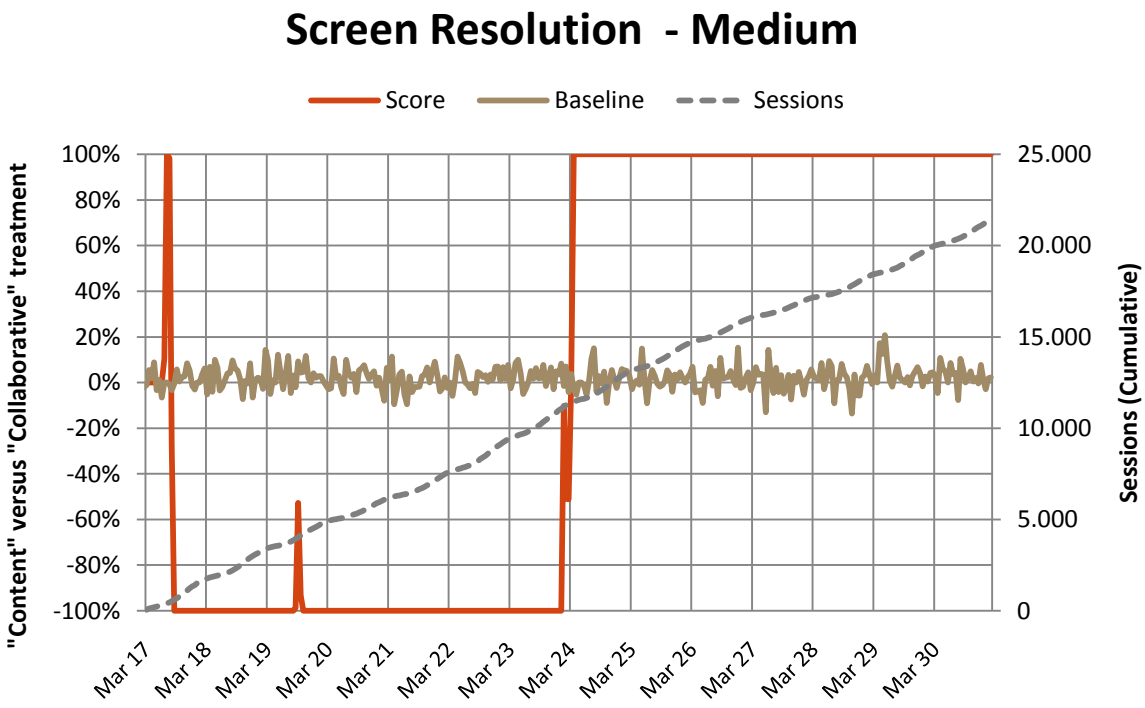


Figure 81: Screen Resolution – Medium

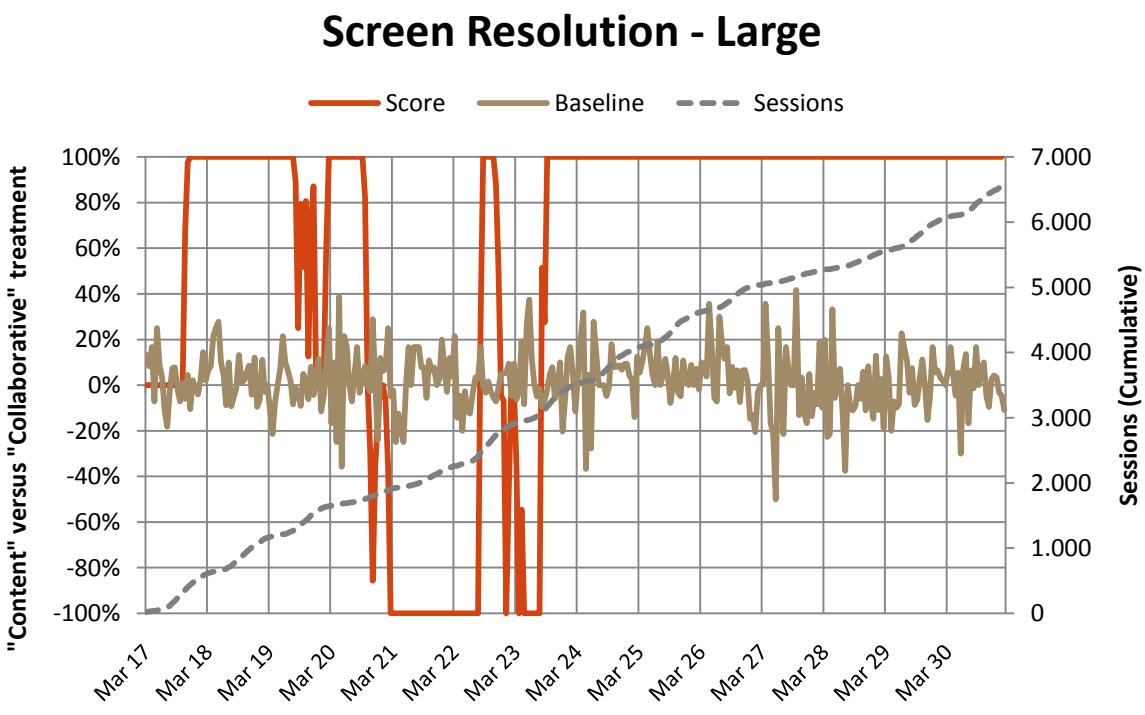


Figure 82: Screen Resolution – Large

K. List of MastersPortal.eu Academic Disciplines

Each Master's programme in the MastersPortal database is linked to at least one of the academic disciplines below. Both the first- and second-level disciplines can be linked. This list of disciplines effectively forms a hierarchy of Master's programmes based upon contextual overlap.

Applied Sciences, Professions & Arts
- Agriculture, Forestry, Animal & Related Sciences
- Design
- Education
- Educational Research
- Ergonomics
- Family and Consumer Science
- Hospitality, Sports, Recreation & Tourism
- Journalism and Mass Communications
- Library and Information Science
- Military Science
- Social Work

Business & Economics
- Accounting
- Business & Technology
- Business Administration
- Econometrics
- Economics
- Entrepreneurship
- Finance
- Human Resource Management
- Management & Organisation
- Marketing
- Project Management
- Public Administration

Engineering & Technology
- Aerospace, Aeronautical & Marine Engineering
- Applied Mathematics
- Bio & Biomedical Engineering
- Chemical Engineering
- Civil Engineering, Architecture & Construction
- Computer Science & IT
- Electrical Engineering
- Energy Engineering
- Engineering & Business
- Engineering Physics
- Environmental & Geo Engineering
- Industrial Design
- Materials Engineering
- Mechanical Engineering

Environmental Sciences
- Climate Studies & Meteorology
- Ecology, Biodiversity & Conservation
- Environmental Biotechnology
- Environmental Chemistry & Toxicology
- Environmental Earth Sciences & Geology
- Environmental Economics
- Environmental Impacts & Human Health
- Environmental Systems Analysis

Environmental Sciences (Continued)
- Environmental Technology
- Geo-information & Spatial Planning
- Hydrology & Water Management
- Soil Science & Soil Ecology
- Sustainable Management, Policy & Governance

Humanities & Art
- Area/Cultural Studies
- Art & Art History
- Film and Theatre Studies
- Language, Literature and Cultural Studies
- Music
- Philosophy
- Religious Studies

Law
- Civil & Private Law
- Criminal Law and Criminology
- European Law
- International Law
- Public Law

Life Sciences, Medicine & Health
- Biomedicine
- Dentistry
- Human Medicine
- Nursery
- Pharmacy
- Physiotherapy
- Public Health
- Veterinary Medicine

Natural Sciences
- Astronomy
- Behavioural Science
- Biology
- Chemistry
- Earth Sciences
- Informatics & Information Science
- Mathematics
- Physics

Social Sciences
- Anthropology
- Communications
- Development & Social Policy and Planning
- Ethnic Studies
- Gender Studies
- Geography
- History
- Linguistics
- Political Science & International Relations
- Psychology

L. Previous Study on Content-Based Recommenders

As part of the coursework for the *Information Retrieval* course (2ID25, autumn 2007) a study on the performance of several content-based recommender systems was executed. The goal of this study was to optimise retrieval of Master's programmes from a large body of unstructured information. This study was executed by Bas Bemelmans[15] and me.
Relevant parts of its results are included *ad verbatim* in this appendix. They serve as a reference for statements made in the *Chapter 2.2.2*. The study is entitled "The Master's Portal Web Miner: Building a Web Mining Application to locate Master's Programmes on University Websites".

L1. Goal of the Study

Gathering sufficient information Master's programmes is a labour intensive task and requires a large investment of money and time. It would therefore be beneficial to be able to automate parts of this process. If we can devise a Web Miner application that searches the Internet for Master's programmes and retrieves information about these programmes, we can save ourselves from this investment.
In this report we will describe how we used three text weighting techniques in an attempt to retrieve relevant Master's programmes from the websites of two Dutch universities.

For our final analysis, we decided to use a larger dataset than we used during previous testing runs. The parameters of our Web Miner were widened, in order to harvest a larger amount of documents. Instead of retrieving a mere 300 documents, as was done during our initial tests, the miner now retrieved in excess of 4.000 documents per university.
The decision to extend the dataset used was based upon the assumption that in most *real-world* scenarios we would encounter bigger, rather than smaller, datasets. Using a bigger dataset will give us a better indication of the real performance of the algorithms. Furthermore, it is important to check whether using this large dataset does not introduce any additional noise.

We harvested information from the websites of the following two universities not currently present in the Master's Portal database:

- University of Maastricht[16]: 5.036 documents
- Radboud University Nijmegen[17]: 4.092 documents

In total 154 megabytes of data was downloaded from the university websites. Mining took about eight hours per university and the speed of mining was mostly restricted by the available CPU-power.

Before we could actually start our analysis, a major issue was discovered in our implementation of the *Pivoted Normalisation* algorithm.
We checked this algorithm on our smaller dataset to see whether it was functioning properly. The initial results were promising as the algorithm seemed to work well. But after we started to use the algorithm on our new, larger, dataset, problems started to occur.
After some searching through our implementation, we concluded that the problems were caused by the logarithmic term in the algorithm. Sometimes, this term generates a so called "not-a-number"

[15] <b.f.n.bemelmans@student.tue.nl>, student number 0534954
[16] http://www.unimaas.nl/
[17] http://www.ru.nl/

output. This obviously occurs when trying to calculate the logarithm of zero or a negative value. Fixing this problem sadly proved to be too much work to complete alongside the writing of this report.
One remarkable issue that we nonetheless would like to note is the fact that for a lot of documents considered relevant by the other two algorithms, *Pivoted Normalisation* returned a relevance score very close to zero. This might indicate that the algorithm does function properly, but that through our problem with the logarithmic function does not provide a correct score.

To discuss the differences between the term weighting algorithms, quantitative output of these techniques was be compared. The comparison will be executed on two different performance indicators: *Quality of the Results* and *Execution Time*.

L2. Quality of the Recommenders

The output that was generated by the *Pivoted Normalisation* algorithm was polluted in such a manner that it could not provide us with useful insights on the quality of results. We therefore decided to only compare the tf-idf and BM-25 weighting schemes in this section

To be able to compare the algorithms in a proper way, we have developed a simple manual scoring scheme to rate the algorithms performance. Our scoring system generates a score for each of the two weighting algorithms. The scores are calculated as follows:

1. Sort the documents by relevancy (c.q. descending) based on the selected weighting algorithm.
2. Find the first 20 documents in the list containing Master's programme information.
    - A Master's programme is defined as the *main* page of the programme. On this page, the general description of the Master's programme should be present.
3. For each Master's programme that is found, its position in the list is added to the algorithms score.
    - For example: If a Master's programme is found in the 13th place of the list, 13 points are added to its score.
4. This process continues until 20 Master's programmes have been identified.
5. The algorithm with the lowest total score is the best performing algorithm.

In an ideal situation, the 20 Master's programmes would be found in the first 20 documents in the list. This way, the lowest possible score of 210 would be achieved. The scoring system was used to calculate the scores for the tf-idf and BM-25 algorithms for both the University of Maastricht, as the University of Nijmegen websites:

- Radboud University Nijmegen
    - tf-idf: 243
    - BM-25: 310
- University of Maastricht
    - tf-idf: 335
    - BM-25: 1012

Looking at these scores, several interesting facts can be concluded. First of all, the tf-idf algorithm outperforms the BM-25 algorithm quality-wise. In other words, the tf-idf algorithm outperforms the

BM-25 algorithm when purely looking at the generated results. Relevant pages are, relatively speaking, rated higher in comparison to BM-25.
These results are somewhat contrary to our initial findings. These findings were done with a much smaller dataset. When using only several hundred documents, BM-25 performs slightly better than tf-idf. This is perhaps because the pages included in the small dataset are much *further* apart. If only very relevant and very irrelevant pages are present, it is more difficult to distinguish between the algorithms.
The score for the University of Nijmegen surely indicates that BM-25 is capable of separating related content from non-related content. After all we should not forget that, out of a set of 4.000 documents, BM-25 manages to place 20 actual Master's programmes within the top 33 results.

A second observation is the big difference in scoring for the two universities. This could be explained by the quality of the actual website in question. The University of Nijmegen has a much higher level of structuring within their website and on the individual pages. Especially this last fact is very important. Pages are very neatly formatted with using a correct HTML-syntax.
For the University of Maastricht, this is not the case. The website is largely unstructured from our Miner's perspective and the HTML documents are very badly formatted and contain a lot repeating elements that are difficult to filter out.
The result of the above can be clearly seen in the results: For the Nijmegen website, the tf-idf algorithm lists 20 pages containing actual Master's programmes in its top 25 results. For the BM-25 algorithm, 20 Master's programmes are found in the top. With some tweaking of the BM-25 parameters, we might be able to increase the performance of the algorithm. In our opinion these results are very promising for our future efforts.

L3. Execution Time of the Recommenders

Apart from the relevancy scores for the three algorithms, the Excel-files also contain the execution times for all of the algorithms. Using this information we can rate the algorithms based upon the time they require to compute their relevancy score.
The execution times for *Pivoted Normalisation* have been included in this analysis. We have no reason to expect that our attempts to fix the algorithm's output will significantly change the execution time of the algorithm.

Pivoted Normalisation is the fastest algorithm, with BM-25 coming in as a close second. The tf-idf algorithm is a distant third. In a way we can say that BM-25 represents the *snowy peaks* on top of Pivoted Normalisation. The average execution times in seconds per document are presented below:

- Radboud University Nijmegen
  - tf-idf: 0,00521
  - BM-25: 0,00117
  - Pivoted Normalisation: 0,00108
- University of Maastricht
  - tf-idf: 0,00625
  - BM-25: 0,00112
  - Pivoted Normalisation: 0,00103

On average, BM-25 is over four times faster than tf-idf. The execution times appear to be a bit longer for documents that receive a high relevance score. This was to be expected, as these documents

have more elements in common with the reference vector. BM-25 and Pivoted Normalisation appear to behave a bit more erratic, having more outliers. Average performance is nonetheless significantly higher.
Interestingly, tf-idf is slower on the University of Maastricht website. As we have already noted, this website has a bad structure and the mined text contains more artefacts. It appears that the document scoring ability of tf-idf it not influenced; only its execution time is. For BM-25 it is not the execution time, but the scoring performance that is affect.

L4. Conclusions

In general, we can conclude that the performance of our Web Miner has exceeded our expectations. For the University of Nijmegen, the Web Miner returns 20 valid Master's programmes in its top 25 of relevant pages. Considering that the Web Miner had a pool of over 4.000 documents to choose from the performance is quite amazing.

Concerning the performance of the individual algorithms, tf-idf seems to provide the best rating quality. The BM-25 algorithm on the other hand might is four times as fast. With some tweaking, it might provide acceptable quality as well.
Our current decision point is thus a trade-off between quality and speed. This decision depends partially on the kind of website we will be mining. Our results show that BM-25 performs almost as well as tf-idf if a website is well structured.