# CAPA: Context Awareness in Predictive Analytics

2011-2015

Project funded by STW under Open Technologieprogramma

**Abstract**. This document contains essential extracts from the accepted STW CAPA project proposal. It is intended to provide an introduction material primarily for applicants for two open positions (4-year PhD and 3-year Postdoc) as well as to the individual researchers and organizations interested in collaboration within the scope of this project.

## TABLE OF CONTENT

## Research Team

| Name | Organization | Role | Expertise |
|---|---|---|---|
| M.Pechenizkiy, dr. | TU Eindhoven | Project leader, PI | Context-awareness (CA) |
| P.M.E. De Bra, prof.dr | TU Eindhoven | AIO Promotor | Adaptive Hypermedia |
| T.G.K. Calders, dr. | TU Eindhoven | Advisor/researcher | Pattern mining |
| I. Žliobaitė, dr. | U. Bournemouth, UK | Advisor/researcher | Concept drift (CD) |
| G. Budziak | Adversitement & TUE | R&D Coordinator | Web Analytics |
| B. Nieme | Adversitement B.V. | Strategy expert | Web Analytics |
| D.M.D. Browne | Adversitement B.V. | Domain expert | Web Analytics |
| P. Lem | Adversitement B.V. | Technical support | Web Analytics |
| E. Verstegen | ReplaceDirect | Domain expert | e-Commerce |
| P. Strizko | MovieMax.eu | Domain expert | Entertainment |
| R. de Winter | Klikniews.nl | Domain expert | Online media |
| T. Putman | studyportals.eu | Domain expert | Education information |
| ***NN*** | TU Eindhoven | ***Open position PhD*** | ***CS, DM/ML/IR or related*** |
| ***NN*** | TU Eindhoven | ***Open position Postdoc*** | ***Data mining, CA, CD*** |

# 1. Project Summary

## 1.1 Research Summary

Web analytics is aimed at understanding behavioral patterns of users of various web-based applications or services: e-commerce, mass-media, and entertainment industries. Within these industries business decisions often rely on two types of predictions: overall or particular user segment demand prediction and individualized recommendations for visitors (prediction of user interests). The main source of data for those predictive analytics tasks is browsing and buying behavior of the visitors. Web analytics application is inherently sensitive to context, which can be defined as a collection of external factors influencing visitor behavior (e.g. location, time, access device, weather, holidays). The behavior may vary depending on the context and potentially within the context. Thus complementing the prediction models with context management mechanisms are expected to make them more specialized and predictive analytics decisions for web applications more accurate.

The project aims to develop a generic framework and corresponding techniques for integrating predictive analytics, context awareness and change detection mechanisms. Scientific novelty, that is in constructing generic approaches for handling the contextual information in predictive analytics, is perfectly aligned with the practical needs. Let us consider an oversimplified example: sales data may indicate increased online spending during the Christmas period. However, without knowing explicitly that this is an important context, such phenomenon is likely to be considered as an anomaly by a predictive model. In general, the number of contextual factors that may potentially affect human behavior on the web is enormous and it is hardly possible to capture all of them with a model simpler than the universe itself. Therefore, one of the key challenges is to construct the mechanisms, which would identify, what the (current) context is and how to integrate it into prediction models. Another important aspect to be developed is the mechanism of monitoring the stream of user-related and contextual data over time to signal anomalies and changes in predictive model performance.

Taking a broad range of practically relevant issues to address within this project, we aim for developing a complete solution allowing straightforward deployment of project results in web-based applications. The techniques we aim to develop will be tested retrospectively on historical data and deployed in real operational settings and validated externally. Research results will be published and lead to writing PhD thesis.

## 1.2 Utilization Summary

The positioning of the project guarantees that the first stage of utilization will start already during the execution of the project. This contrasts with the current state in the field of web analytics, where research and development are rather fragmented and often considered a commercial secret that prevents the dissemination of the current state of the art. In fact, the research community typically has access to extracts of real historical data at best and opportunities to deploy and test the methods in real settings are very limited. This project will facilitate such opportunities for the project team.

The research and deployment will undergo several cycles, where in the first cycle, the current state of the art approaches will be adopted and validated in *laboratory* and *field* experiments, and in the following cycles more advanced issues of context awareness and change detection integration into predictive web analytics will be continuously developed and validated.

Three main participating end users will be the companies owning the websites ReplaceDirect.nl, MovieMax.eu and Kliknieuws.nl. The technology user Adversitement B.V. will facilitate the infrastructure for diverse business case studies with the end users (e-commerce, entertainment and media). ReplaceDirect.nl is a client of Adversitement, but participates in the project as an independent partner. The developed techniques will be validated externally for alternative utilization cases: aggregated demand prediction and individual recommendations. The particularly planned prediction tasks include but are not limited to bidding for sponsored search, content based and collaborative recommendations, demand prediction for a given time period, and content matching in online advertising.

The participating users expressed their interest to directly deploy the results in the web analytics systems they develop and/or use. Their motivation is reflected in the letters of support and contribution. However, the expected utilization is by no means restricted to the end user companies and the technology company involved in the project. The generic framework and techniques we aim to develop will benefit a wide range of users representing multiple industries. Yet, we believe that continuous deployment and validation of the research results in real operational settings, that is planned to take place within the scope of this project, will guarantee a high-quality, trustable research output, which is then much more likely and easy to be utilized by the interested parties.

# 2. Scientific description

We are looking at web based applications, where user behavior is the main source of data. Understanding behavioral patterns has been recognized as an essential component of web analytics. Web analytics is the measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage. Accurately predicting the probability of desired actions on the web in specific circumstances would enable us to achieve personalization and adaptation to diverse customer needs and preferences. Prediction tasks, relevant for business decisions, can be split into aggregated demand prediction and individualized recommendations for visitors.

The desired action depending on the business objectives can take many forms. Examples include sales of products, membership registrations, newsletter subscriptions, software downloads, or any activity beyond simple page browsing. The interests and behavior of a visitor might vary depending on the context and potentially within the context (e.g. location, time, access device). Context may be temporal, geographic, based on activity or behavior. Informally, context characterizes the circumstances the entity is in, a setting.

Consider a video on demand site (TV shows, movies, documentary). End user behavior on the site might be different depending on the context of her need. A user might search for a specific documentary needed for reference while at *work* or might be browsing newly released movies while at *home*. Purchase behavior might vary depending on a *holiday* or the *weather outside*, which are examples of contextual factors.

Predictions in web analytics are inherently context sensitive. Web analytics is aimed at analyzing the behavior of internet users, i.e. humans in all their varieties, diversities, inconsistencies. A number of contextual factors that may affect visitor behavior on the web is enormous and it is hardly possible to model the behavior explicitly. We aim to take context and changes into predictive models for web analytics applications. In addition, web marketing provides a natural ground for integration of multichannel data, where a channel can be viewed as a dimension of context.

## 2.1 Research contents/Introduction

The proposed project aims to develop new techniques and tools for business intelligence, particularly the generic framework and corresponding techniques for integrating predictive analytics, context awareness and change detection mechanisms. This will allow managing marketing budgets more efficiently and effectively, and reducing the chances that the customers are exposed to undesired or irrelevant content.

Context awareness is needed in predictive web analytics, since the circumstances under which decisions are made are not static. It would allow integrating external explanatory information into the learning process, aiming to reduce uncertainty for the learning models. Integrated change detection mechanisms would inform about unexpected behavior to reduce the chances of misleading marketing actions.

### a. *Research questions*

The main goal of the project is to develop a generic framework for designing context-aware prediction techniques for web analytics. We formulate the following research question: **How can we integrate context awareness and change detection into predictive web analytics in order to achieve better user(s) behavior prediction accuracy?**

To be able to integrate context awareness into predictive modeling we have to address the following subquestions (that will be detailed further in the time plan and division of tasks section):

1. How to define the context (form and maintain contextual categories) in web analytics?
2. How to connect context with the prediction process in predictive web analytics?
3. How to integrate change detection mechanisms into the prediction process in web analytics?
4. How to ensure integration and feedback mechanisms between change detection and context-awareness mechanisms?
5. What should a reference architecture allowing to plug in new context aware prediction techniques for a collection of web analytics tasks look like?

### b. *Research intent and CAPA model prototype*

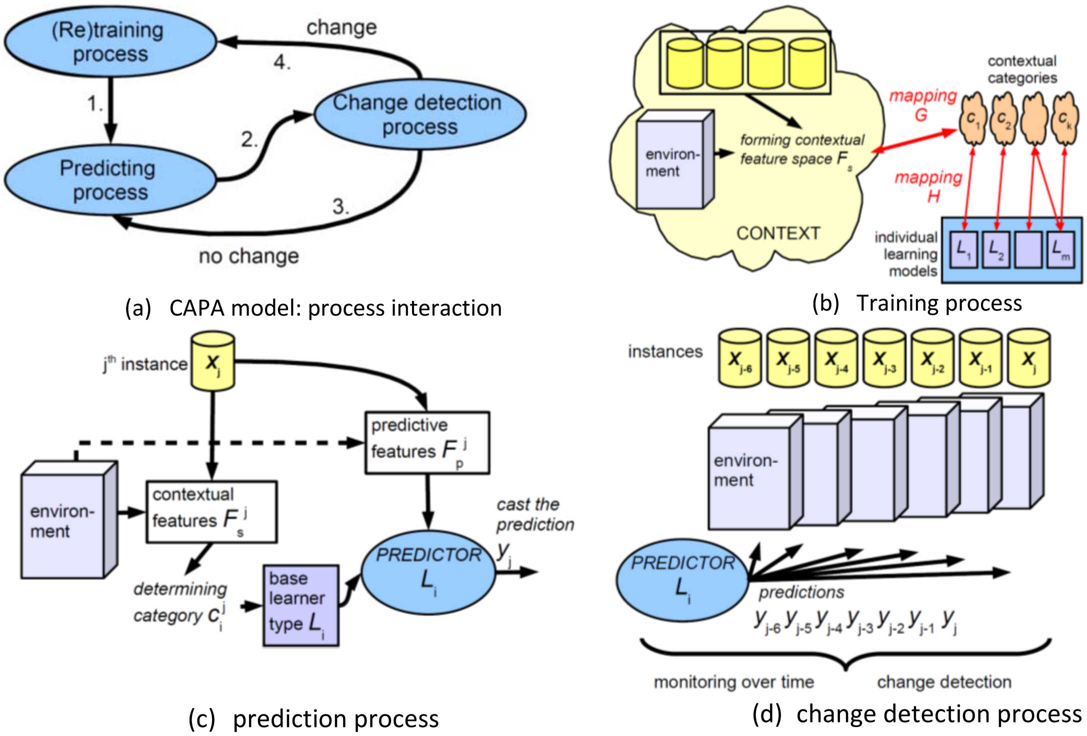
**Data overview.** Web analytics is centered on collecting, processing and analyzing click-stream data. Click-stream data can be aggregated, grouped at different levels of granularity and integrated with customer data, content data and external contextual data (e.g. calendar holidays, weather, media highlights) to formulate various supervised learning tasks.

The reporting categories typically include internet traffic data (page views, visits, unique/returning visitors, click-paths), navigation related data (link clicks, traversed paths), origin data (referrer, search engine, keywords), system data (browser, resolution, plug-ins), and commercial and purchase related data.

**Conceptual reference model.** A prototype of the conceptual reference model for a context aware prediction system equipped with change detection mechanisms consists of three main processes: *(re)training*, *prediction* and *change detection* (Figure 1a). First model design choices need to be made. Then the model can be trained using historical data (*training process*). After that unseen instances are sequentially received, their contexts are determined and predictions are output (*prediction process*). Regularly the historical data sequences and performance of the model are inspected to identify possible changes (*change detection process*). If a change is detected, the model is retrained.

Assume a website presenting sports news in chronological order. The publisher is interested to predict the average relevance of an individual news item at a given time. Assume that the relevance of news varies for different reader groups depending on the geographic location, or the type of access device. A temporary or permanent change in relevance patterns might occur. For example, due to the World Championship in Football a number of readers might shift to Africa thus temporarily forming a new geographical context. The introduction of a new revolutionary product (e.g. iPad) may permanently change reading patterns.

In this example, the *training process* forms prediction rules using historical data. *The prediction process* applies these rules to predict the relevance of a news item. The c*hange detection process* monitors the performance of the predictor to notice changes (a) within context, (b) of context, or (c) of current contextual categories or mapping. Once a change is detected, the system goes to the *(re)training process* to adapt the prediction rule, update the contextual categories, or alert the necessity of redesigning the model.



(a) CAPA model: process interaction

(b) Training process

(c) prediction process

(d) change detection process

**Figure 1. Prototype of the generic context-aware prediction model with change detection.**

**(Re)Training process (Figure 1(b)).** Consider a supervised learning set up. Let $X \in \Re^p$ be an object of interest with a label $y \in Z$, for example $X$ is a football news article, $y$ is the number of visitors for this article. The ultimate task is to learn a label prediction function $y = L(X)$, what will be the relevance (cumulative interest) of a particular article in particular context, e.g. relevance of an article describing an injury of Dutch national team member within the next hour?

We integrate context awareness into the design of the prediction system to restrict the space of search of $L$ types and parameterizations. For that contextual features $F_s$ are constructed, which are describing the environment of an instance $X$. For example, it may matter whether today is a holiday or working day, an average season or the World Cup. Let $C = (c_1, c_2, ..., c_k)$ be a set of contextual categories, a procedure for constructing them is defined by a designer. Let $G:F_s \to C$ be a mapping from contextual features to contextual categories. Let $L = (L_1, L_2, ..., L_m)$ be a set of individual learning procedures (defining the selection of training instances, input feature space, classification technique and its parameterization) or already learnt models. The two key ingredients of a context-aware learning design are: defining the contextual features $F_s$ with a mapping $G$, and fixing the mapping $H:C \to L$. For example, if it is a holiday, we use one prediction, if it is a working day, we use different prediction rule for sports news items.

**Prediction process (Figure 1(c)).** Two level decision making for a given object $X_j$ has two steps:

1. the individual predictor $L_i$ (or a combination) is selected $L_i = H(G(C))$,
2. the decision is made using $L_i$ as the predictor $y_j = L_i(X_j)$.

First the contextual category of a new instance is determined; then the corresponding predictor is applied to this new instance. The first level can be seen either as a switch mechanism from one context to another at an instance level or generally as a context identification mechanism.

**Change detection process (Figure 1(d)).** The change detection process includes monitoring the data input and model output at an instance subset level. It complements the context switch mechanism. Let $t$ indicates time stamp 'now'. Then $X_1, X_2, ..., X_t$ is a sequence of instances up to 'now', $F_{s1}, F_{s2}, ..., F_{st}$ is a sequence of contextual features, $y_1, y_2, ...y_t$ is a sequence of predictor outputs. These three sequences will be analyzed to signal outliers and changes. While the predicting process operates at the level of a single instance, the change detection process considers the group (set or sequence) of instances.

The role of the change detection process is to trigger model updating if necessary. We require change detection to be able to distinguish changes from outliers (temporary) and not trigger updating due to outliers. For example, outliers like a website being down or disruption in data transmission mess the statistics and might need to be removed from the historical data when learning the predictors. A change like volcano eruption disturbing flights might alter online airline ticket sales temporarily (more browsing, less purchasing) or even persistently (purchasing closer to the departure date due to uncertainty).

Integrated change detection would call for three levels of actions depending on the type and severity of a change. First, retraining only the predictors; second, updating the contextual categories followed by retraining the predictors; third, reconsider model design. The first two are considered to be a part of CAPA model. The third would give an alert to the analyst to reconsider some of the design choices.

**Model design.** Note that the three processes do not include model design. Model design means fixing the elements of the model and the way it has to be (re)trained. By training we mean a particular parameterization (e.g., choosing a linear regression as a predictor is a design choice, while learning its parameters from historical data refers to the model training process). In the CAPA model the design choices to be made concern: (1) integration of data sources, (2) possible contextual and predictive features, (3) possible contextual categories and categorization method, (4) choice of individual predictors.

**Foreseen challenges.** It is not always known beforehand, which context is useful to consider, or which features describe the context. In a basic scenario one can explicitly assume that a mapping of contextual categories to the learnt classifiers is known. Consider a sports news example, let the aim be to predict the relevance of a given news item in the next hour. Higher relevance of all football news can be assumed during the time of a championship. A designer can directly introduce two models based on these contexts: one for ordinary days, one for championships. It is different from including a championship feature into predictive feature space, as model splitting allows different types of predictors, different feature selection.

In more complex cases both the contextual relevance and the mapping to the predictors are unknown. A procedure needs to be defined for forming contextual categories and mapping them to the predictors. For example, based on the domain knowledge and/or historical data we can identify two contextual categories for the same user profile: searching the news ‘for work' and 'for leisure'. We can learn to separate these based on time of the day, browsing patterns, other contextual features we can find or construct from contextual data.

### c. *Research Contribution*

**Intended research results.** The main goal of the project is to develop a framework and a collection of techniques how to integrate context awareness and change detection into predictive web analytics.

Starting from the prototype presented in Figure 1 we will study algorithmic aspects and analyze the performance of the two level decision making for alternative web analytics application scenarios. We will develop a generic framework including techniques for forming contextual categories and for linking them with the predictors for integrating context awareness and change detection into predictive models. We will also produce a set of guidelines for using the proposed framework for designing new techniques.

The techniques will be tested retrospectively on historical data as well as deployed and validated online in the field experiments. Taking a broad approach in terms of relevant research content we aim for a complete solution that would allow the deployment in web analytics.

**Scientific importance and novelty.** Predictive web analytics is a rather distributed research field due to various commercial interests involved. Increasing volumes of data and web related commercial activities provide opportunities to model and predict user needs more precisely. However, the loads of data also pose a challenge to data mining due to different types of dynamics and external effects in the behavior.

Even more commercial activities are expected to shift to the web in the near future. The web has the advantage of round the clock opening hours, location independent services, reduced personnel costs and thus lower prices for the end customers, but most importantly – they provide personalized and tailored services. Thus adaptivity of the prediction systems and context awareness is crucial to accommodate customer needs.

Along with these tendencies and the complexity of influential factors, massive learning techniques are not enough; models need to be tailored to take contexts into account. In general marketing the information is gathered to segment the client population and as a result to better serve the needs of each specific segment. The same holds for predictive web analytics: specific models are needed for each context.

Scientific novelty is in the generic approach for handling the contextual information, integration of context awareness and change detection. In web analytics this is of primary importance due to direct relation to customer segmentation and profiling. In addition, web marketing provides a natural ground for integration of multichannel data, where a channel in principle can be viewed as a known context.

### d. *Research methodology*

This project involves basic and applied research aimed to benefit from each other. The multi methodological approach (conceptual-theoretical, constructive and experimental) will be adopted.

In the *conceptual-theoretical approach*, conceptual basics and formalisms of the generic CAPA will be developed. First, a taxonomy of context-aware approaches will be built. Then the applicability and limitations of the existing techniques w.r.t. the properties of the real application problems will be identified. In the *constructive approach* the developed techniques will be embedded into the prototype CAPA system to test it through the *experimentation approach* and to facilitate the subsequent refinement of the theory and the prototype in an iterative manner. Testing of research ideas is intended to be done using MATLAB software and/or the MOA [27] research environment. System prototyping and integration into real operational setting for external validation will be done with the use of the operational infrastructure provided by Adversitement BV.

The traditional experimental data mining research paradigm will be used for the **internal evaluation** of the developed framework and corresponding techniques on a set of reference and online real-world datasets (some of which are already made available for the project). Progressive evaluation (time-wise) and cross validation (object-wise) procedures will be employed. **External validation** of our work will be performed through the integration of the developed techniques into web analytics systems (including, but not limited to O2MC maintained by Adversitement, which will be described in more detail in the next section).

We will employ traditional A/B and multivariate testing procedures providing reliable estimates of the performance of the alternative approaches. We have the commitment from the participating companies to conduct such procedures. The thorough continuous evaluation process consisting of internal and external validation procedures is an essential part of both theory building and testing. Phasing of the project is presented in Section 2.3. Graphical representation of one research cycle is provided in Appendix A.

## 2.2 Existing infrastructure

The research will physically take place at TUe (Eindhoven) and partly also at Adversitement BV (in Uden and Amsterdam) that will stand as intermediary for transforming the research results into application tailored to different online business models. Adversitement will provide support for the case studies facilitating continuous deployment and external validation of developed framework and techniques in real operational settings.

Adversitement runs an intelligent web analytics system platform (O2MC). Currently the platform is primary used for forming ad campaigns and bidding in sponsored search advertisements. It evaluates the quality of visitor actions for every individual keyword and automatically increases/decreases the cost per click when necessary. O2MC is meant to operate as a generic recommendation platform. It is positioned as an intelligent decision support tool, which combines the data and delivers real time marketing decisions.

This research benefits two types of companies: end users (web sites and web shops) and technological users (web analytics services). The end user companies either have internal personnel for maintaining web analytics or outsource this service to consulting companies like Adversitement. The research output is not in any way restricted to be used in interaction of both types of users. The presence of a technological user in the project first makes the project timing more efficient, via facilitating the infrastructure for testing. Secondly, close interaction with a technological user during the course of the research project will catalyze utilization of the project results.

## 2.3 Time plan and division of tasks

As stated in the research subquestions (Section2.1.a), the project research content consists of four main elements: (1) defining the contexts, (2) integrating contexts into prediction models, (3) integrating change detection into prediction models, and (4) utilization of the context aware models in web analytics tasks. The research elements will be executed iteratively and repeatedly for three case studies. The planned project timing and the interaction between the PhD and postdoc research lines are depicted in Figure 2.

We expect to hire a postdoc having adequate background in change detection, learning under concept drift. Background in context aware learning would be an advantage. The research line of PhD student position tightly follows designing the framework and tools for context aware learning. The postdoc will have two research lines - change detection and context awareness.

**Year 1.** The PhD student will start by studying context awareness and related fields, in order to acquire an inventory of relevant methods and techniques. Then (s)he will define the context part of the project, addressing the first research question “How to define the context?”. It includes the following subquestions: “Which contextual features are relevant to consider and how to obtain them?", “How to handle possible overlap between contextual and predictive features?", “How to integrate contextual features associated with a single instance and the ones which can be obtained from a set (e.g. increasing temperature)?".

The postdoc (assumed to have a background in change detection) starts with research experiments, adapting and implementing the current state-of-the-art change detection techniques (e.g. [26, 9, 55]). To facilitate the repetitive research cycles of immediate internal and further external validation of the developed framework and approaches, the prototyping will go side by side with theory building.

The joint work between the PhD student and postdoc will focus on developing and testing the basic CAPA model, integrating context awareness and change detection into predictive web analytics. The basic set up assumes that contextual features and categories are known, and only sudden changes can take place.

In **Year 2** the PhD student will investigate the cases when contextual features are unknown and contextual categories are to be formed. Temporal context and contextual clustering approaches will be explored [54] and alternative techniques to infer the contexts will be introduced. The PhD student will address the subquestion: “How to integrate context if it is not directly measurable as such?".

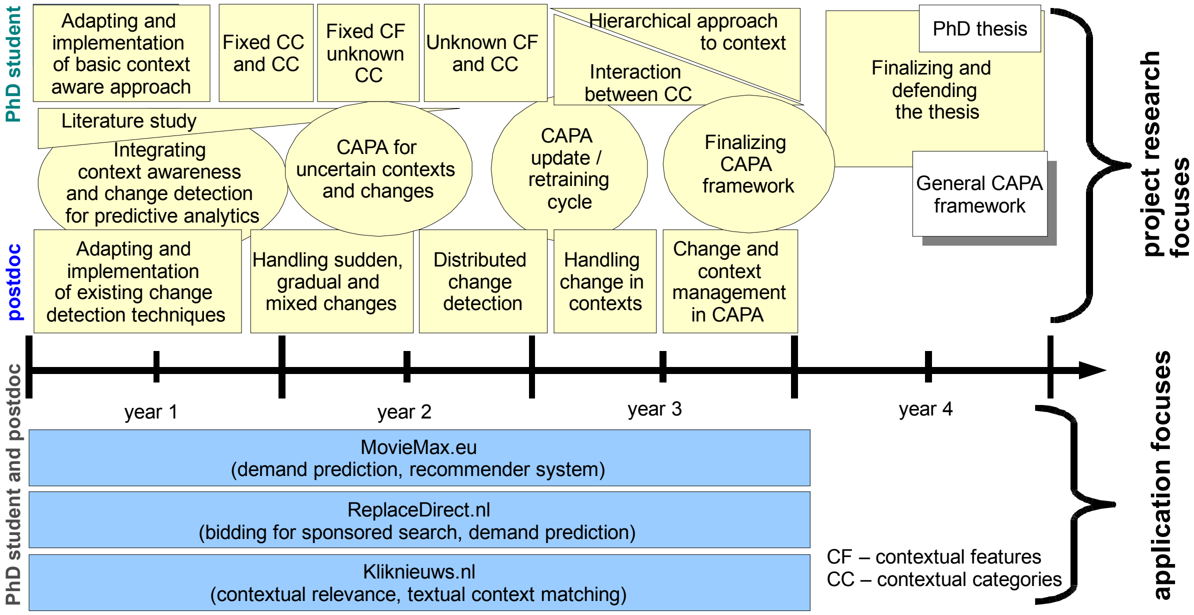

PhD student
Adapting and implementation of basic context aware approach
Fixed CC and CC
Fixed CF unknown CC
Unknown CF and CC
Hierarchical approach to context
Interaction between CC
PhD thesis
Finalizing and defending the thesis
Literature study
Integrating context awareness and change detection for predictive analytics
CAPA for uncertain contexts and changes
CAPA update / retraining cycle
Finalizing CAPA framework
General CAPA framework
project research focuses
postdoc
Adapting and implementation of existing change detection techniques
Handling sudden, gradual and mixed changes
Distributed change detection
Handling change in contexts
Change and context management in CAPA
year 1
year 2
year 3
year 4
PhD student and postdoc
MovieMax.eu
(demand prediction, recommender system)
ReplaceDirect.nl
(bidding for sponsored search, demand prediction)
Kliknieuws.nl
(contextual relevance, textual context matching)
CF – contextual features
CC – contextual categories
application focuses

**Figure 2: Project timeline: working approaches and resulting outputs**

The postdoc will integrate change detection techniques, suitable for different change types (sudden, gradual) and develop distributed change detection techniques, which are particularly relevant in context aware web analytics, since the changes might be observed in multiple products, contexts, channels.

With the selected elements, a sketch of the general conceptual framework will be created. The main goal of year 2 research activities will be to research how to connect contextual and change detection information with the prediction process in cases of known and unknown context.

By **Year 3** the PhD student and postdoc are expected to be advanced with the basic setup of the model and will address more complex contexts and change cases, as listed in the timeline (Figure 2). The main goal of the third year will be to finalize the general framework, which will be complemented by a set of techniques to handle basic and complex contexts and detect changes of mixed types.

In **Year 4 the** PhD student will work on final evaluation of the developed model and finalizing the thesis.

The user committee will observe the research and continuously provide feedback about the case studies.

**Outcomes.** The research will result in a generic framework for context-aware predictive analytics, with a set of externally validated techniques. The range of practically relevant issues researched in this project will lead to a complete solution ready for wide adoption in industries for which predictive web analytics matters. Research results will be published in peer-reviewed venues and lead to writing one PhD thesis.

## 3. Utilization plan

The project will induce a close interaction of scientific research, technology user and the end users. Close cooperation with the technology user facilitates faster utilization of the results. The first stage of utilization will start already during the execution of the project, when the developed techniques will be deployed in online testing. Based on the (successful) cases we expect the research results to spread widely to other end users and possibly to other industries, especially since they will be extensively validated online.

Currently the practical deployment of data mining techniques in web analytics is analogous to mass marketing. The contexts in which decisions are made and the environmental dynamics are not a part of the computational decision support models. Context tailored models would make predictions more precise.

We are going to develop a generic framework and a set of supporting techniques to utilize the contextual information in web analytics. The research and deployment will undergo three main cycles, where first state of the art will be adopted, context awareness and change detection will be integrated into predictive web analytics, a model prototype will be developed and tested in laboratory and field experiments, see Appendix A. This way utilization of the results will start already during execution of the project.

Within the scope of the project at least three extensive business case studies will be carried out. We are going to address aggregated (demand) and personal prediction (recommendations) tasks. Each selected combination of a task and model advancement will undergo the research-deployment cycle presented in Appendix A. The intended case studies are:

- movie demand prediction and video on demand recommender system with the **MovieMax.eu** case,
- keyword performance in a sponsored search prediction and demand prediction for business customers with the **ReplaceDirect.nl** case,
- news recommendation and online content matching using the **Kliknieuws.nl** case.

These cases represent different e-industries (e-commerce, entertainment and information services). In the three types the desired visitor actions and key performance indicators (KPI) are different. However, context awareness in web analytics is relevant to all cases. Potential users are by no means limited to these case studies. The presence of a technology user catalyzes dissemination of the results.

The three case studies cover a wide range of prediction applications in web analytics, namely demand prediction for resource planning and management of technical capacities, recommender systems for online and offline recommendations of items, determining a bid price for sponsored search advertising, online textual content matching, and others. The intended case studies are detailed in Appendix B.

The case studies were chosen to represent the main prediction tasks encountered in web analytics. The listed users will provide data and domain expertise for case studies. The technological user will facilitate the field experiments. Thus utilization of the results will start already during the course of the project.

**Utilization and Impact. Field testing will be performed** on three distinct online business cases. The users will give feedback on the ongoing research while attending the user committee meetings. A series of case studies will be published to disseminate the findings to the research community and industry. The generic framework will benefit a wide scope of users representing multiple e-industries. The presence of a technology user gives the project the competitive advantage of having a field testing environment and domain expertise.

As indicated in the attached letters of support, the users will be eager to directly deploy the results in their web analytics systems. Naturally, the deployment is not binding, but they are motivated to utilize the technologies to be developed and thus are participating. In addition to the supporting users everyone who is interested will have open access to the results published in scientific reports, papers and case studies.

# 4. Positioning of the project proposal

The project has exceptional potential for utilization due to the actively involved technological user Adversitement BV. We have no knowledge of any other initiative similar to the proposed project.

Related research can be grouped into three types of relations: (1) similar vocabulary ('contextual learning', 'context awareness') and philosophy, (2) similar techniques and methods, designed for solving different problems, (3) applications, addressing intelligent methods for web analytics.

**Context awareness** is broadly exploited in different fields such as ubiquitous computing, however, taking a different focus while sharing the vocabulary with the proposed project. Most of the works assume that the context is given. It is not addressed how to formulate context and categories. Context awareness has been analyzed mainly from an application building point of view. However, there is a lack of analysis from an algorithmic point of view, which is our focus. There have been research results making the first steps (rather conceptual contributions) how to discover contexts from the data [54, 61].

**Related machine learning and data mining** approaches can be found on the crossroads of context awareness [54, 61], meta learning [13, 57, 48, 44, 4], multiple classifier systems [37] and concept drift [59, 53, 24, 17, 30, 38, 31] research fields. There is a broad body of methods for change and anomaly detection, see [7, 56]. The specifics of change detection in supervised learning are summarized in [23]. A part of concept drift research has concentrated on change detection and reaction to it, including publications by the research team [32, 41]. A few works have tried to identify and reuse reoccurring contexts [33, 25, 69]. Yet, we are not aware of the works directly addressing a combination of learnable contexts and change detection for unfamiliar situations. Recent works in contextual bandits [63, 43] come close from a problem formulation perspective, while concentrate on reinforcement learning.

**Related web analytics research** includes work on change detection in visitor behavior, mostly focused on retail customers, typically mining association rules. Song et al [49] detect three types of changes using association rule mining: emerging patterns, unexpected changes and added/perished rules. [65] predicts visitor behavior in internet shopping malls using association rule mining to determine purchasing patterns and then to form a feature space for predicting purchasing probability. Recommenders [3] have been exploited in e-commerce sites. Works on contextual advertisement [18, 40] refer to positioning of an ad in relation to the viewed content and therefore come conceptually close to CAPA framework.

**From an application perspective**, there are a number of vendors proposing web analytics solutions, e.g. HBX Analytics or Google Analytics. Although each tool has its own features and characteristics, their focus is directed towards data integration and reporting; and not algorithmic prediction solutions.

An expanded overview of the related research is presented in Appendix C.

## 4.1 Uniqueness of the proposed project

The research in web analytics is fragmented; besides it is often considered a commercial secret. That inhibits dissemination of the current state of the art. The research community typically has access only to extracts of historical data. Opportunities to test the methods in real settings are limited, but are essential in this dynamic field. Simulation on historical data is prone to unintentional overfitting, while the true data for 'what if' is not available. This project will be build around prototyping opportunities, allowing continuous field testing. This will allow a number of feedback iterations during the framework development.

From a research perspective, the project aims to integrate context awareness and change detection into predictive web analytics which is a unique direction itself. This project, in contrast to the mainstream concept drift research, goes beyond the temporal context. Intuitive contexts (time, location, mood, activity) do not necessarily have a direct correspondence to contextual categories. In contrast to context-awareness research (e.g. ubiquitous computing) we consider both 'meaningful' and 'abstract' contexts.

Close interaction with the participating companies within planned case studies, makes deployment and utilization directly expected. Moreover, the case study companies will participate in the user committee. Thus in this project the user committee will have a uniquely active role to input the domain knowledge and directly monitor and influence the ongoing research.

### 4.2 Embedding of the proposed project

The project provokes collaboration within the Information Systems group headed by prof. De Bra. The relevant core competencies of the group include data mining, adaptation and personalization.

Three NWO projects in progress within the group are related. The Handling Concept Drift in Adaptive Information Systems (HaCDAIS, led by dr. Pechenizkiy) is developing a unifying framework and data mining techniques for handling concept drift, which can be regarded as a change in context. The Generic Adaptation Framework (GAF, led by prof. dr. De Bra) researches the design and implementation of AIS, aiming at analysis and definition of a new reference model. It relates via evolution towards automatic detection of changes in user behavior and adaptation. The Complex Patterns in Streams (COMPAS, led by dr. Calders) is focused on the algorithmic aspects of mining sequential patterns and evolving graph patterns in data streams. The relation to CAPA is via the data stream perspective.

## List of publications cited

[1] A. Achilleos, K. Yang, and N. Georgalas. Context modeling and a context-aware framework for pervasive service creation: A model-driven approach. Pervasive Mob. Comput., 6(2):281-296, 2010.

[2] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems, pages 335-336, New York, NY, USA, 2008. ACM.

[3] A. Albadvi and M. Shahbazi. Integrating rating-based collaborative filtering with customer lifetime value: New product recommendation technique. Intelligent Data Analysis, 14(1):143-155, 2010.

[4] A. Tsymbal and M. Pechenizkiy and P. Cunningham and S. Puuronen. Dynamic integration of classifiers for handling concept drift. Information Fusion, 9(1):56-68, 2008.

[5] A. Tsymbal and M. Pechenizkiy and P. Cunningham and S. Puuronen. Dynamic integration of classifiers for handling concept drift. Information Fusion, Special Issue on Applications of Ensemble Methods, 9(1):56-68, 2008.

[6] L. Baltrunas and F. Ricci. Context-based splitting of item ratings in collaborative filtering. In RecSys '09: Proc. of the 3rd ACM conference on Recommender systems, pages 245-248, 2009.

[7] M. Basseville and I. V. Nikiforov. Detection of abrupt changes: theory and application. Prentice-Hall, Inc., 1993.

[8] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni. A survey of context modelling and reasoning techniques. Pervasive Mob. Comput., 6(2):161-180, 2010.

[9] A. Bifet and R. Gavalda. Learning from time-changing data with adaptive windowing. In SIAM Int. Conference on Data Mining, 2007.

[10] D. Billsus and M. J. Pazzani. User modeling for adaptive news access. User Modeling and User-Adapted Interaction, 10(2-3):147-180, 2000.

[11] C. Bolchini, C. Curino, E. Quintarelli, F. Schreiber, and L. Tanca. A data-oriented survey of context models. SIGMOD Rec., 36(4):19-26, 2007.

[12] M. Bottcher, M. Spott, D. Nauck, and R. Kruse. Mining changing customer segments in dynamic markets. Expert Syst. Appl., 36(1):155-164, 2009.

[13] P. Brazdil, J. Gama, and B. Henery. Characterizing the applicability of classification algorithms using meta-level learning. In F. Bergadano and L. D. Raedt, editors, ECML, volume 784 of LNCS, pages 83-102. Springer, 1994.

[14] N. Bricon-Souf and C. Newman. Context awareness in health care: A review. International Journal of Medical Informatics, 76(1):2-12, 2007.

[15] J. Burrell, G. K. Gay, K. Kubo, and N. Farina. Context-aware computing: A test case. In UbiComp '02: Proceedings of the 4th international conference on Ubiquitous Computing, pages 1-15, London, UK, 2002. Springer-Verlag.

[16] H. Cao, D. Hu, D. Shen, D. Jiang, J. Sun, E. Chen, and Q. Yang. Context-aware query classification. In SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 3-10, 2009.

[17] G. Castillo and J. Gama. Adaptive bayesian network classifiers. Intell. Data Anal., 13(1):39-59, 2009.

[18] D. Chakrabarti, D. Agarwal, and V. Josifovski. Contextual advertising by combining relevance with click feedback. In WWW '08: Proc. of the 17th int. conf. on World Wide Web, pages 417-426, 2008.

[19] H. Chang, Y. Tai, and J. Hsu. Context aware taxi demand hotspots prediction. Int. J. Bus. Intell. Data Min., 5(1):3-18, 2010.

[20] N. Chen, A. L. Chiu, and H. H. Chang. Mining changes in customer behavior in retail marketing. Expert Syst. Appl., 28(4):773-781, 2005.

[21] M. Davis, M. Smith, J. Canny, N. Good, S. King, and R. Janakiraman. Towards context-aware face recognition. In Multimedia'05: Proceedings of the 13th annual ACM international conference on Multimedia, pages 483-486, New York, NY, USA, 2005. ACM.

[22] P. Dourish. Seeking a foundation for context-aware computing. Hum.-Comput. Interact., 16(2):229-241, 2001.

[23] A. Dries and U. Ruckert. Adaptive concept drift detection. Statistical Analysis and Data Mining, 2(5-6):311-327, 2009.

[24] J. Gama and G. Castillo. Learning with local drift detection. In X. Li, O. R. Zaïane, and Z. Li, editors, ADMA, volume 4093 of LNCS, pages 42-55. Springer, 2006.

[25] J. Gama and P. Kosina. Tracking recurring concepts with meta-learners. In L. S. Lopes, N. Lau, P. Mariano, and L. M. Rocha, editors, EPIA, volume 5816 of LNCS, pages 423-434. Springer, 2009.

[26] J. Gama, P. Medas, G. Castillo, and P. P. Rodrigues. Learning with drift detection. In Advances in Artificial Intelligence - SBIA 2004, 17th Brazilian Symposium on Artificial Intelligence, proceedings, volume 3171 of Lecture Notes in Computer Science, pages 286-295. Springer, 2004.

[27] G. Holmes, A. Bifet. Moa massive online analysis. http://www.cs.waikato.ac.nz/~abifet/MOA/

[28] P. D. Haghighi, A. B. Zaslavsky, S. Krishnaswamy, M. M. Gaber, and S. W. Loke. Context-aware adaptive data stream mining. Intell. Data Anal., 13(3):423-434, 2009.

[29] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster. The anatomy of a context-aware application. Wirel. Netw., 8(2/3):187-197, 2002.

[30] E. Ikonomovska, J. Gama, R. Sebastião, and D. Gjorgjevik. Regression trees from data streams with drift detection. In J. Gama, V. S. Costa, A. M. Jorge, and P. Brazdil, editors, Discovery Science, volume 5808 of LNCS, pages 121-135. Springer, 2009.

**[31] I. Žliobaitė. Combining similarity in time and space for training set formation under concept drift. Intelligent Data Analysis, 15(4):in press, 2011.**

**[32] J. Bakker and M. Pechenizkiy and I. Žliobaitė and A. Ivannikov and T. Karkkainen. Handling outliers and concept drift in online mass flow prediction in CFB boilers. In Proceedings of the 3rd International Workshop on Knowledge Discovery from Sensor Data (SensorKDD'09), pages 13-22. ACM Press, 2009.**

[33] I. Katakis, G. Tsoumakas and I. Vlahavas. Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowledge and Information Systems, 22(3):371-391, 2010.

[34] J. Kjeldskov and M. Skov. Exploring context-awareness for ubiquitous computing in the healthcare domain. Personal Ubiquitous Computing, 11(7):549-562, 2007.

[35] L. Kovács, P. Mátételki, and B. Pataki. Service-oriented context-aware framework. In Proc. of the 4th European Young Researchers Workshop on Service Oriented Computing, volume 2 of EPTCS, pages 15-26, 2009.

[36] A. Kreutzmann, K. Terzic, and B. Neumann. Context-aware classification for incremental scene interpretation. In UCVP '09: Proceedings of the Workshop on Use of Context in Vision Processing, pages 1-6, 2009.

[37] L. I. Kuncheva. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004.

**[38] L. I. Kuncheva and I. Žliobaitė. On the window size for classification in changing environments. Intelligent Data Analysis, 13(6):861-872, 2009.**

[39] S. Lawrence. Context in web search. IEEE Data Eng. Bull., 23(3):25-32, 2000.

[40] W. Li, X. Wang, R. Zhang, Y. Cui, J. Mao, and R. Jin. Exploitation and exploration in a performance based contextual advertising system. In Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2010.

**[41] M. Pechenizkiy and J. Bakker and I. Zliobaite and A. Ivannikov and T. Karkkainen. Online mass flow prediction in cfb boilers with explicit detection of sudden concept drift. ACM SIGKDD Explorations Newsletter, 11(2):109-116, 2010.**

[42] M. A. Munoz, M. Rodríguez, J. Favela, A. I. Martinez-Garcia, and V. M. González. Context-aware mobile communication in hospitals. Computer, 36(9):38-46, 2003.

[43] N. Pavlidis, D. Tasoulis, and D. Hand. Simulation studies of multi-armed bandits with covariates. In UKSIM '08: Proc. of the Tenth International Conference on Computer Modeling and Simulation, pages 493-498. IEEE Computer Society, 2008.

**[44] M. Pechenizkiy, A. Tsymbal, S. Puuronen, and D. W. Patterson. Feature extraction for dynamic integration of classifiers. Fundam. Inform., 77(3):243-275, 2007.**

[45] R. Reichle, M. Wagner, M. U. Khan, K. Geihs, J. Lorenzo, M. Valla, C. Fra, N. Paspallis, and G. A. Papadopoulos. A comprehensive context modeling framework for pervasive computing systems. In 8th IFIP WG 6.1 Int. Conf. on Distributed Applications and Interoperable Systems (DAIS), volume 5022 of LNCS. Springer Verlag, 2008.

[46] J. Scanlan and J. Hartnett. A context aware scan detection system. IJCSNS International Journal of Computer Science and Network Security, 8(1):75-84, 2008.

[47] B. N. Schilit, N. Adams, and R. Want. Context-aware computing applications. In In Proceedings of the Workshop on Mobile Computing Systems and Applications, pages 85-90. IEEE Computer Society, 1994.

[48] C. Soares and P. Brazdil. Selecting parameters of svm using meta-learning and kernel matrix-based meta-features. In H. Haddad, editor, SAC, pages 564-568. ACM, 2006.

[49] H. S. Song, J. K. Kimb, and S. H. Kima. Mining the change of customer behavior in an internet shopping mall. Expert Systems with Applications, 21(3):157-168, 2001.

[50] T. Strang and C. Linnhoff-Popien. A context modeling survey. In Proc. of the Workshop on Advanced Context Modeling, Reasoning and Management as part of the 6th Int. Conf. on Ubiquitous Computing (UbiComp 2004) , 2004.

[51] E. Suh, S. Lim, H. Hwang, and S. Y. Kim. A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study. Expert Syst. Appl., 27(2):245-255, 2004.

[52] K. Torkkola, M. Gardner, C. Schreiner, K. Zhang, B. Leivian, H. Zhang, and J. Summers. Understanding driving activity using ensemble methods. In D. V. Prokhorov, editor, Computational Intelligence in Automotive Applications, volume 132 of Studies in Computational Intelligence, pages 39-58. Springer, 2008.

[53] A. Tsymbal. The problem of concept drift: Definitions and related work. Technical report, Department of Computer Science, Trinity College Dublin, 2004.

[54] P. Turney. The management of context-sensitive features: A review of strategies. In Proc. of the ICML-96 Workshop on Learning in Context-Sensitive Domains, pages 60-65, 1996.

[55] M. van Leeuwen and A. Siebes. Streamkrimp: Detecting change in data streams. In Proc. of Machine Learning and Knowledge Discovery in Databases, European Conference (ECML/PKDD 2008), Part I , volume 5211 of Lecture Notes in Computer Science, pages 672-687. Springer, 2008.

[56] A. B. Varun Chandola and V. Kumar. Anomaly detection: A survey. ACM Computing Surveys, 41(3):article 15, 2009.

[57] R. Vilalta and Y. Drissi. A perspective view and survey of meta-learning. Artificial Intelligence Review, 18(2):77-95, 2002.

[58] Z. Wen, M. Zhou, and V. Aggarwal. Context-aware, adaptive information retrieval for investigative tasks. In IUI '07: Proc. of the 12th international conference on Intelligent user interfaces, pages 122-131, 2007.

[59] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. Machine Learning, 23(1):69-101, 1996.

[60] J. Yuan and Y. Wu. Context-aware clustering. In Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, pages 1-8, 2008.

[61] M. Zacarias, H. S. Pinto, and J. Tribolet. Automatic discovery of personal action contexts. In HCP-2008 Proc., Part II, MRC 2008 - 5th Int. Workshop on Modeling and Reasoning in Context, pages 75-88. TELECOM Bretagne, 2008.

[62] K. Zhang, K. Torkkola, H. Li, C. Schreiner, H. Zhang, R. M. Gardner, and Z. Zhao. A context aware automatic traffic notification system for cell phones. In ICDCS Workshops, page 48. IEEE Computer Society, 2007.

[63] S. Zhong, A. Martinez, T. Nielsen, and H. Langseth. Towards a more expressive model for dynamic classification, Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), 2010.

[64] P. Brezillon, Ch. Tijus: Modeling user search on a website by contextual graphs. Revue d'Intelligence Artificielle 23(4): 467-484, 2009.

[65] E. Suh, S. Lim, H. Wang and S. Y. Kim. A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study Expert Syst. Appl., 27: 245-255, 2004.

[66] M. Antonie, and O. R. Zaïane, An associative classifier based on positive and negative rules. In Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. DMKD '04. ACM, New York, NY, 64-69, 2004.

[67] K. Zettsu, and K. Tanaka, Referential Context Mining: Discovering Viewpoints from the Web. In Proceedings of the 2005 IEEE/WIC/ACM international Conference on Web intelligence. Web Intelligence. IEEE Computer Society, 321-325, 2005.

[68] A. Sen, P.A. Dacin, and C. Pattichis, Current trends in web data analysis. Communications of ACM 49, 11, 85-91, 2006.

[69] J.B. Gomes, E. Menasalvas, P.A.C. Sousa. Tracking recurrent concepts using context. In Proceedings of Rough Sets and Current Trends in Computing, LNAI 6086, pp. 168-177, 2010.

[70] M. Spiliopoulou and L. Faulstich. WUM: A tool for Web utilization analysis. In Proceedings of 6th Int. Conf. on Extending Database Technology (EDBT'98), Valencia, Spain, March 1998.

[71] B. Berendt and M. Spiliopoulou: Analysis of Navigation Behavior in Web Sites Integrating Multiple Information Systems, The VLDB Journal, Vol. 9, No. 1, pp. 56–75, May 2000.

## Definitions

**Conversion rate** - The percentage of visitors who take a desired action.

**Technological user** – user of technology to be developed in the course of the project.

**Web marketing** - marketing activities to acquire customer to online stores and retain them.

**Web analytics** is the measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage.

**Predictive analytics** encompasses a variety of techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events.

**Predictive web analytics** aims to predict individual and aggregated characteristics indicating visitor behavior for purposes of understanding and optimizing web usage.

**Web mining** is the application of data mining techniques to discover patterns from the web, including web usage mining, web content mining and web structure mining.

**An outlier** is an observation that lies an abnormal distance from other values in a random sample from a population. This definition leaves it up to the analyst to decide what will be considered abnormal. It is defining a single instance.

**An anomaly** is any occurrence or object that is strange, unusual, or unique. It can also mean a discrepancy or deviation from an established rule or trend. It is not supposed to last, but can be several instances in a sequence.

**A change** is a transition from one state of the process to another. If anomaly lasts for long it is a change.

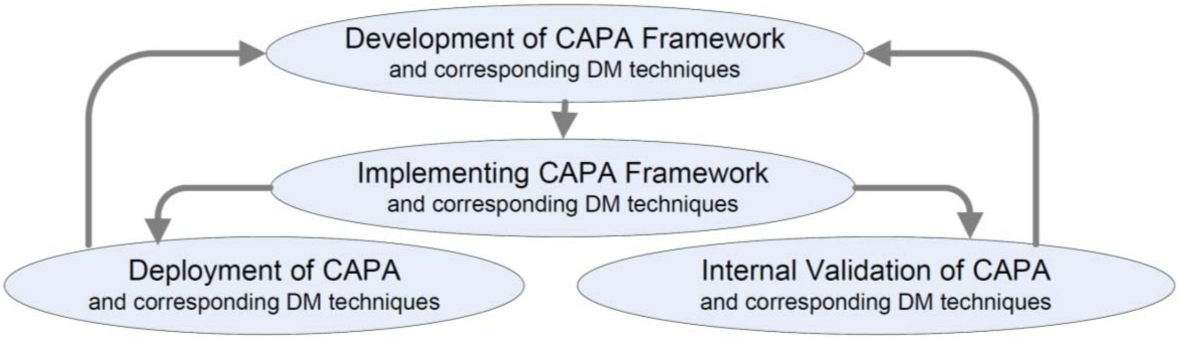
**A drift** is a form of change, typically referring to gradual transition.

**Context** characterizes the circumstances the entity is in, a setting.

**Context switch** term typically is used in parallel computing, but we use it in a different meaning. By context switch we mean a change in contextual category (usually in a timeline). For example a change from summer to winter.

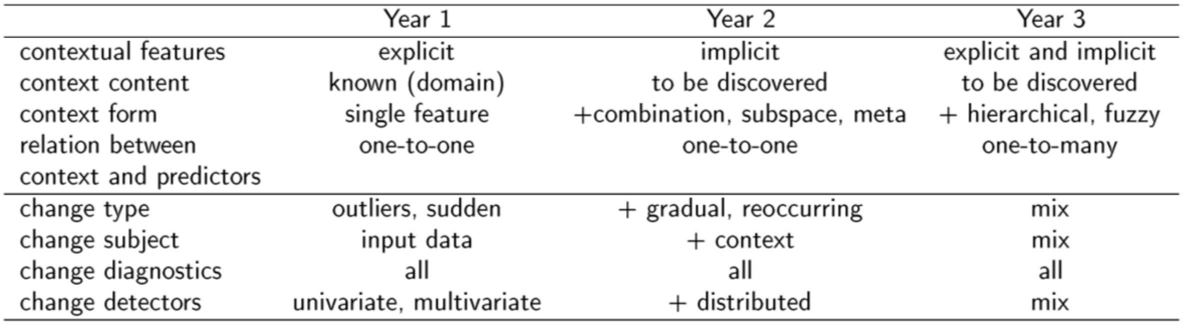
**Contextual advertising** is advertising on a Web site that is targeted to the specific individual who is visiting the Web site. A contextual ad system scans the text of a Web site for keywords and returns ads to the Web page based on what the user is viewing, either through ads placed on the page or pop-up ads.

## Appendix A. R&D cycles following research methodology


Development of CAPA Framework
and corresponding DM techniques
Implementing CAPA Framework
and corresponding DM techniques
Deployment of CAPA
and corresponding DM techniques
Internal Validation of CAPA
and corresponding DM techniques

From context awareness and change detection perspective, the main research issues can be summarized with respect to the several categories. The corresponding research content to be covered in the three years of the project is summarized in Table 1.

**Table 1. Scenarios integrating context awareness and change detection**



| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| contextual features | explicit | implicit | explicit and implicit |
| context content | known (domain) | to be discovered | to be discovered |
| context form | single feature | +combination, subspace, meta | + hierarchical, fuzzy |
| relation between context and predictors | one-to-one | one-to-one | one-to-many |
| change type | outliers, sudden | + gradual, reoccurring | mix |
| change subject | input data | + context | mix |
| change diagnostics | all | all | all |
| change detectors | univariate, multivariate | + distributed | mix |

# Appendix B. Intended Case Studies

## Movie Max

MovieMax.eu (Entertainment Retail Group BV) is a movie retailer also providing video on demand (VOD) services. VOD systems either stream content through a set-top box, a computer or other device, allowing viewing in real time, or download it to a device such as a computer or portable media player for viewing at any time. Lately, the relative share of VOD in movie rental is rapidly increasing. The video on demand market is expected to grow in the Benelux in the coming three-four years as well as access to the VOD equipment. The studies within Movie Max will address VOD application.

Two case studies will be carried out. The first will address the problem of total demand prediction for given time slots and for given products. Context awareness and change detection is directly relevant firstly because of the streaming nature of the task. Some contextual features are indicated to be directly relevant, based on the company domain knowledge. For instance, the relation of the demand to weather.

The results of this case study will contribute to more efficient resource management. First of all, from a technological perspective there are restrictions and limitations on bandwidth usage, thus accurate demand prediction would allow to plan accordingly, including diversified pricing. Furthermore, longer term demand prediction would allow to advance in negotiating the rates with video content and network bandwidth suppliers.

The available (historical) data includes 2-3 years of video on demand online sales history, specifying date, customer, movie, genre. A domain based movie categorization is available.

The second case study will concentrate on recommender systems for video on demand. In recommender systems the role of context has been acknowledged [2] and is intuitively explicable. For instance, a different movie might be preferred depending on the social context (alone, with friends, with girlfriend). Two types of recommendation tasks will be addressed online and offline.

To facilitate online testing three technical ingredients will be needed: multivariate (AB testing) facilities, label collection through review and implicit review collection systems (information about watching time, breaks via server logs).

Overall, the integration of context awareness and change detection is directly relevant to VOD application. A part of these tasks are currently being handled based on human expertise and heuristic rules of thumb.

## Replace Direct

ReplaceDirect is an e-commerce site selling spare parts for personal computers. Visitor actions on the website are concentrated around orders, which are directly related to the company revenues. The ultimate desired action is obviously ’checkout’.

The case study within ReplaceDirect will primary address sponsored search advertising. Large part of sales come from visitors directed from the search engines, both sponsored and organic search. It is essential to optimize the cost/benefit trade-off in the sponsored search. Thus for each keyword we will aim to predict the conversion rate (desired actions) for the upcoming time period based on the context, in order to optimize bidding prices. In addition, we are interested to detect sudden changes in behavior in order to trigger the update of bidding prices and long term trends in order to retrain the prediction model on time.

Currently available data for the bidding case contains historical performance of several hundred keywords on a daily basis for the period of 2 months, more is expected to be accumulated by the start of the project. The available data includes pre-click and post-click attributes. Pre-click attributes are related to the interaction with the search engine, and include number of impressions, number of clicks, maximum and average CPC (cost per click), average ad position. Post-click attributes are sequences of visitor actions on the site, including desired actions. Contextual features from external sources are to be added, for instance, time, holidays, weather, interest rates.

Using this data an offline lab experiment will be conducted starting with the 10-20 largest volume keywords aiming to learn the next bidding price. After satisfactory performance offline is achieved, A/B testing will be implemented to asses the performance online.

In addition to the bidding task, we intend to carry out follow up case studies within ReplaceDirect in relation to CAPA. The discussed tasks include demand prediction for stock management (which is subject to contextual factors) and recommender system for business customers (repeated purchases in higher volumes) which would be relevant in offline marketing.

### Kliknieuws

Kliknieuws.nl is an online news portal. It provides a stream of regional or categorized news. The news text is produced within the company (it is not news aggregation). Currently news items are listed in chronological order. Basic contextual matching is applied for ad placement.

We are going to address two tasks within this case study.

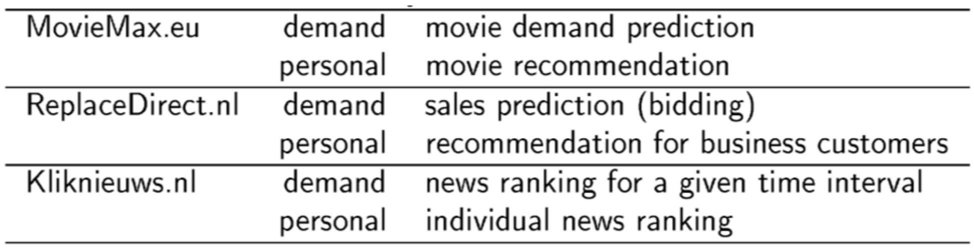
The first task is content based news recommendation for a reader. We expect to especially benefit from contextual factors in this case; since the news is regional thus specific contextual factors might be localized well. Holidays, local events (e.g. market day), event traffic jams are expected to contribute to the relevance prediction accuracy, especially since the news portal is very regional. Based on the observed context the relevance of current news items will be reassessed and presented accordingly.

Context aware recommendations are expected to be particularly relevant when reading from mobile devices, which are gaining popularity. Screen size and typical use time there do not allow reading everything and even to overview everything at on sight.

The second task will address contextual news access and contextual advertisement tasks in this study. The second task is content based add matching. Here add relevance depends not only on the textual content, but also on the external factors (or context). Again, for example, users might have different interests in given ads depending whether it is holiday and weekday, even though content relation to the news item might be the same.

An important issue to be addressed in relation to advertisement is negative context. We need to identify not only the most relevant, but also avoid textual matching in negative context. For example, children clothing is not to be advertised in line with pedophilia scandal, although content match is perfect, the context is not right.

**Table 2. Summary of the intended case studies**



| | | |
|---|---|---|
| MovieMax.eu | demand | movie demand prediction |
| | personal | movie recommendation |
| ReplaceDirect.nl | demand | sales prediction (bidding) |
| | personal | recommendation for business customers |
| Kliknieuws.nl | demand | news ranking for a given time interval |
| | personal | individual news ranking |

# Appendix C. Related Research

We give a broad overview of related research initiatives to provide a full check and mapping of the CAPA project. In this Appendix we provide an expanded discussion of context awareness research and machine learning/ data mining techniques, which are related via using similar vocabulary and similar philosophy behind the techniques, but typically designed for different problems.

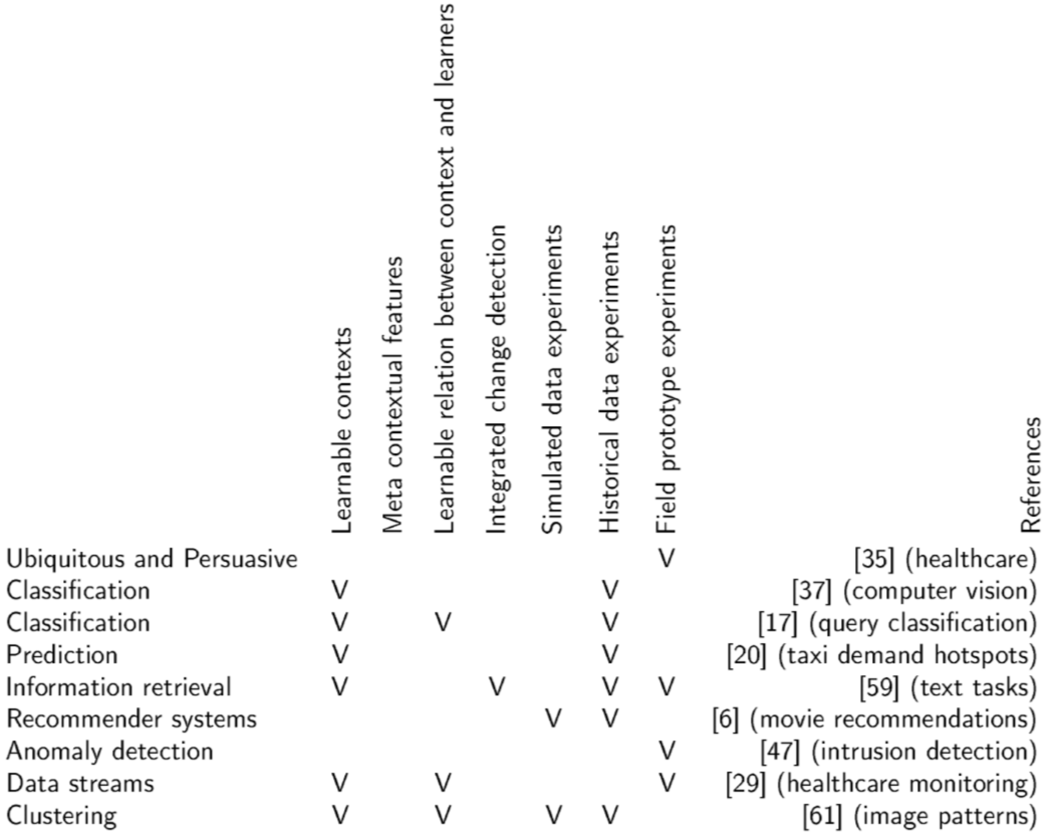
## Related Context Awareness Research

Context awareness as a concept has been promoted initially in ubiquitous computing, where geographical context, user context (personal), and device context play important roles. General application frameworks have been developed [47, 29]. Then context awareness spread to other areas including machine learning (e.g. automotive systems [52, 62], face recognition [21]), information retrieval (e.g. contextual web search [39, 64]) and recommender systems [10, 2], mobile computing in general [22], health care [14], mobile communications in hospital [42], mobile tour guides [15].

In the past decade a number of generic frameworks for context modeling and reasoning have been developed in pervasive/ubiquitous computing [50, 8, 11, 45, 35, 1], which typically consider only physical pre-specified contexts (time, geographical location), which are not being learned within the training process. These frameworks typically address software and middleware architecture issues rather than algorithmic learning aspects.

In Table D1 we present recent examples of studies related to context awareness over different research areas and applications. In spite of very diverse use of the term 'context-aware' it commonly refers to model splitting based on some fixed or learnable characteristics in order to make the models more specialized to particular circumstances. In general the setups are very diverse and distant from supervised learning notion, mainly investigating implementation of model localization mechanisms assuming more or less given context.

**Table 3. Context awareness across research areas.**



| | Learnable contexts | Meta contextual features | Learnable relation between context and learners | Integrated change detection | Simulated data experiments | Historical data experiments | Field prototype experiments | References |
|---|---|---|---|---|---|---|---|---|
| Ubiquitous and Persuasive | | | | | | | V | [35] (healthcare) |
| Classification | V | | | | | V | | [37] (computer vision) |
| Classification | V | | V | | | V | | [17] (query classification) |
| Prediction | V | | | | | V | | [20] (taxi demand hotspots) |
| Information retrieval | V | | | V | | V | V | [59] (text tasks) |
| Recommender systems | | | | | V | V | | [6] (movie recommendations) |
| Anomaly detection | | | | | | | V | [47] (intrusion detection) |
| Data streams | V | | V | | | | V | [29] (healthcare monitoring) |
| Clustering | V | | V | | V | V | | [61] (image patterns) |

## Related machine learning and data mining approaches

From a research perspective the proposed research lays primarily on the crossroads of context awareness, meta learning, multiple classifier systems and concept drift research fields.

Meta learning uses meta data about the machine learning experiments to categorize the performance of the learning algorithms and thus facilitate intelligent algorithm selection given (the properties of) a dataset at consideration [13, 57].

In the CAPA model (Figure 1), finding a mapping between contextual categories and individual predictors can be performed by meta learning. In existing works meta-learning is applied mostly either at the dataset level [48] or at the instance level [44, 4]. In the proposed research there will be a need to consider different scenarios where the context has to be derived for the currently observed single instance (or a set of instances) rather than a complete dataset. Besides, contextual features can be formed not only from the data describing instances but also from external sources of information (e.g. weather, holidays).

Multiple classifier systems (MCS) allow simultaneous use of arbitrary feature descriptors and classification procedures [37]. CAPA model (Figure 1) relates to MCS via maintaining a set of individual predictors. In principle, MCS as a paradigm is rather generic and can be (and will be) used to develop the framework.

Concept drift is often referred to as changes in hidden contexts or underlying data distribution affecting supervised and unsupervised learning [59, 53]. Recent years have been very productive in the research community bringing a number of adaptive learning algorithms [24, 17, 30], including publications by the research team [5, 38, 31].

There is a broad body of methods for change and anomaly detection, see [7, 56]. The specifics of change detection in supervised learning are summarized in [23]. A part of concept drift research has concentrated on change detection and reaction to it, including publications by the research team [32, 41].

Relatively few works have tried to identify and reuse reoccurring contexts [33, 25].We are not aware of the works directly addressing a combination of learnable contexts and change detection for unfamiliar situations.

Recent advances in contextual bandits [63, 43] come close from a problem formulation perspective. In the literature, contextual bandits are sometimes called bandits with covariate, bandits with side information, associative bandits, and associative reinforcement learning. The bandit approach differs from supervised learning formulation (our formulation) in a way that in supervised learning the values of all actions are known (historical training data), while in reinforcement learning exportation takes place when action-reward knowledge comes in an iterative manner.

Related web analytics applications. The research concerning web analytics and web marketing applications is rather fragmented. This research now is mainly driven by data availability. Web marketing applications oriented research typically concerns e-commerce websites, internet shopping malls in particular. Typically these applications are driven by web mining research that can be roughly divided into web content mining, web structure mining and web usage mining. Web usage mining has the most direct relation to advancing the analysis of navigation behavior in web sites [70], possibly integrating multiple information systems [71].

Previous studies that apply web usage mining to internet shopping malls in [51] are divided into three groups: basic algorithm, web site redesign, recommendation. A typical use of web usage mining here is reduced to association rules mining to determine purchasing patterns and then use those purchasing patterns as features, i.e. to form an attribute space for predicting purchasing probability. Mining associative classifiers [66] can be also utilized with a similar goal. Albadvi and Shahbazi [3] apply collaborative filtering for recommendations in e-commerce sites. There is a body of literature on change detection in visitor behavior, mostly focused on retail customers, typically mining association rules. Song et al [49] detect three types of changes using association rule mining: emerging patterns, unexpected changes and added/perished rules. Bottcher et al [12] analyze the change of frequent itemsets over time to detect changes in customer behavior in physical shopping. Chen et al [20] aim to detect changes in customer behavior in retail marketing using behavioral, demographic variables and a transaction database. Detecting changes in the context - external environmental variables is absent in the field.

Works on contextual advertisement [18, 40] refer to positioning of an ad in relation to the viewed content and therefore come conceptually close to CAPA framework. The current use of word context here has different meaning and is not defined explicitly. They do not address external context for prediction, but rather refer to positioning of an ad in relation to the viewed content. We are talking about varying user preferences depending on the context. Besides, the existing approaches are ad hoc, i.e. they typically use a predefined match between keywords occurring in the viewed content and desired ad placement.

However, use of web content mining and text mining approaches has a clear potential for this type of applications.

Web structure mining approaches have not been used for predictive analytics. However, similarly to the works in contextual advertisement, in this area a few approaches have been suggested to compute contextual scores, e.g. reputation, of web pages. That is, instead of computing a global plain (e.g. page rank) score, hyperlink (anchor text) and its surrounding content can be used to derive the context (see so-called referential context mining approach [67] for an example).

From an application perspective, there are a number of vendors proposing web analytics solutions: SiteCatalyst, Omniture, Google Analytics, DoubleClick, NedStat, Webtrends. Although each tool has its own features and characteristics, the data collection architecture is similar (in so far as it has been published). Historically web analytics is used to monitor, analyze and report visitor behavior on sites leaving extrapolation to the future and prediction decisions to human experts (marketing). Apparently the attention is directed towards data integration and reporting, algorithmic prediction solutions are typically very limited, and tailored to specific applications. The statement "Current surveys suggest that in spite of storing many terabytes of clickstream data and investing heavily in Web analytic tools, few companies understand how to use the data effectively." [68] made a few years ago is still describing the current state-of the art in web analytics fairly well.