



Master projects, internships and other opportunities

Utrecht, February 2011



1. Table of Contents

1. Table of Contents	2
2. About Teezir.....	3
3. Automatically detect emerging topics and trends.....	3
4. Contextual disambiguation for search queries.....	3
5. List completion / automatic benchmarking	3
6. Multilingual Opinion Mining.....	4
7. Spelling correction for search.....	4
8. Information extraction from webpages.....	4
9. High performance Javascript parsing.....	4
10. Data structures for incremental crawling.....	4
11. Improving sentiment analysis.....	5
12. Helping users make successful queries	5
13. Tracking online content	5
14. Contact.....	5

2. About Teezir

Teezir is a young and innovative technology company that develops and deploys comprehensive search solutions. Teezir lets companies take advantage of large and diverse amounts of documents or texts, using break through search technology.

Teezir's search platform provides functionality for the entire process of disclosing data: from gathering content, analyzing documents and building indexes for efficient access to effective querying and ranking of information. Teezir's framework is based on full-text retrieval techniques.

The projects described in this document aim to add value to our Opinion Mining and Focused Content solutions. Teezir's Opinion Mining technology collects and analyses web based user generated content, and thus enables companies to gain insight into online customer opinions, topics and trends. Teezir's Focused Content solution aims to find, collect and unlock highly targeted content about a specific topic or for a specific niche market.

Visit our website at <http://www.teezir.com> to learn more about Teezir, or try our opinion mining technology at <http://www.WhoRules.nl>.

3. Automatically detect emerging topics and trends

For brand monitoring, campaign evaluation and other marketing related activities, knowing what people think or say about products is pivotal. Teezir's opinion mining solution uses advanced search technology to aggregate views and opinions found on the web, in discussion groups or blogs, and to create statistics about what people are saying. Querying this data allows decision makers to slice and dice the content, and learn what people say, either at the very aggregated level: "what is the share of positive versus negative views about our new product?", or at the very detailed level: "which sources reflect this negative sentiment, and what exactly are people saying?".

To create an overview of what is said in the context of, for example, a given brand, topic or theme, Teezir currently identifies the terms that are distinguishing in this context. We are looking to create novel ways of given end users more insight in the dynamics of topics and trends. Which novel topics and trends are emerging? Where does the buzz start? How do topics evolve over time?

4. Contextual disambiguation for search queries

Search requests are short. A typical user types at most 2 or 3 terms in a search box. With this little information it can be hard to find the real information need or intention behind a search request. Recently, many studies have looked at using contextual information to disambiguate queries.

In many of Teezir's opinion mining solutions users can pose multiple queries at once. This is perhaps most apparent in our online showcase <http://www.WhoRules.nl> where users are encouraged to directly compare the sentiments on a number of topics. This context of other queries is a valuable source for disambiguating search requests. For example the term *brown* probably has a different meanings in the comparisons *brown vs blair* and *brown vs green*. Similarly, in the Dutch comparisons *bos vs balkenende* and *bos vs strand* we are probably looking for sentiments on very different meanings of *bos*. This project aims improving search results by taking such query contexts into account.

5. List completion / automatic benchmarking

In Teezir's opinion mining solutions users can compare results on different queries. This way our clients can for example benchmark sentiments w.r.t. their own brand against the competition. Similarly, in our online opinion mining showcase users can directly measure and compare the sentiment on multiple topics. While we do suggest related topics in our more extended opinion mining solutions, currently users need to manually define and type their own comparisons.

In this project we aim to develop algorithms for automatically completing a list of compared items as well as for suggesting comparisons or benchmarks on a given topic. For example, when someone compares *ada*, *pnda*, *vvd*, the

system could suggest to also add *d66*, *pvv*, *sp* and *groenlinks*. In the second phase of the project we could look at automatically producing such lists given a general topic. For example, for the topic *soccer* the system might suggest *ajax*, *psv*, *feenoord*, *fc twente*, *az*, *nac*

6. Multilingual Opinion Mining

Teezir's Opinion Mining solution (eCare) collects analyzes and summarizes user-generated content from a multitude of blogs, forums and review sites. Our focus has been predominantly on Dutch and English data and we are working on sentiment analysis for other European languages. So far, we have treated each of these languages in isolation. This separate summarization of results for individual languages creates the opportunity to highlight differences in sentiments, topics and trends between languages. While that information is very useful, it is also important to produce language independent overview statistics.

In this project we will study the application of cross language information retrieval (CLIR) techniques to opinion mining. Can we directly combine sentiment scores from different languages? Can existing CLIR techniques be directly applied for topic analysis to find meaningful term clouds?

7. Spelling correction for search

Working with data collected from the web means dealing with spelling errors. Because of the abundance of web data, spelling errors are not necessarily an issue, since any spelling variant is bound to exist somewhere and yield results. But is this also the case for opinion mining? Do we miss important sentiments when spelling variants are not taken into account?

This project aims to improving our spelling correction mechanisms for opinion mining. Do standard spelling correction algorithms apply or do we need special techniques to guarantee optimal sentiment analysis? What is the impact of applying spelling correction on opinion mining result quality?

8. Information extraction from webpages

The ultimate goal of any search system is to exactly point the user to the information that was requested. The current search systems do quite a good job on presenting relevant pages and snippets, the addition of an information extraction component would make it possible to extract entities that are listed on webpages, such as the people, products, or prices. This makes it possible for the user to pose queries like: "Give me the cheapest toothpaste."

The challenge is to extract these entities without writing site specific scripts. This challenge is hard; however, we expect that using the structure of webpages (such as table-like structures) will help a lot to deliver the right information. This work builds upon the Master Thesis of Samuel Louvan [1]

9. High performance Javascript parsing

In the past years, Teezir has worked on creating a web crawler capable of handling Javascript. While this web crawler works fine, there's still much research and development that can be done to improve the performance. One of the main challenges is handling Javascript in a high performance way.

Javascript is usually invoked as an interpreted language that makes it possible to change the DOM tree of a web page. During the implementation of Javascript into our web crawler, we noticed a lot of Javascript patterns that resulted in similar behavior. We believe it's possible to automatically learn or otherwise use the behavior of commonly used patterns to create a very high performance javascript engine for web crawlers.

10. Data structures for incremental crawling

Downloading a couple of web pages and storing their content is the easy part of web crawling; keeping terabytes of information as up to date as possible is the hard part. During web crawls, a lot of different information is stored by a web crawler. Some pages are cacheable, some are not, some are candidate for indexing and some contain

interesting metadata. This makes storage in a database unsuitable. However, for updates a normal database is very suitable.

The combination of these aspects is what moved companies like Google to a technology called BigTable, which has different semantics than standard databases. We believe that BigTable is only the first step towards fulfilling the actual requirements. A second step would be to tune the data structure of the tables for optimal scheduling, while keeping the performance requirements intact.

11. Improving sentiment analysis

Teezir's Opinion Mining solution (eCare) collects analyzes and summarizes user-generated content from a multitude of blogs, forums, microblogs, and review sites.

Teezir is continuously testing and improving its sentiment analysis. The goal of this assignment is to discover and test new directions of improving sentiment analysis. For example, using metadata (such as the source, date, query or author), improved negation detection, or improving our wordbanks. Teezir will provide you with (test-) collections. Furthermore, you can base your work on previous work done at Teezir [2].

12. Helping users make successful queries

Teezir's Opinion Mining solution (eCare) collects analyzes and summarizes user-generated content from a multitude of blogs, forums, microblogs, and review sites. One of the unique features of eCare is that it stores historical data together with free-querying. This means that a user can pose any possible query to eCare to view the sentiment and other analysis around a topic.

Possible queries could be brands (e.g. "Ziggo" or "ABN AMRO") or products (such as "books" or "cable television"). For these types of queries it is often the case that people will literally use the query terms in their posts if they write about the subject (e.g. "I like the cable television offer of Ziggo."). However queries are also used to describe more complex concepts such as "quality" or "price". In this case it is more complicated to create a query that matches posts that write about the subject.

The goal of this assignment is to design a computer aided method to help users to make successful queries for complex concepts. This could include statistical analysis of the data set, using example postings, a standardized brain-storm process etc. etc.

13. Tracking online content

Teezir's Opinion Mining solutions (eCare) monitors the online volume and sentiments around specified search requests. This assignment would extend this to monitoring the flow of specific documents on the internet, to answer two types of questions. First, monitoring of the origin of a topic or discussion: Where is it first posted? Second, monitoring what happens with the content posted by the user (e.g., a viral campaign that was started): Who's picking it up? Where does it appear next?

Techniques to follow the flow of content on the internet could take from existing technology for Query by Example search, Plagiarism detection, and Link Analysis.

14. Contact

If you are interested in doing a project with Teezir, please contact us at info@teezir.com.

[1] Web Page Segmentation & Structure Analysis for Eliminating Nonrelevant Content by Samuel Louvan
August 2009

[2] Analyzing Sentiment in a Large Set of Web Data while Accounting for Negation by Bas Heerschop, Paul van Iterson, Alexander Hogenboom, Flavius Frasinca, and Uzay Kaymak, to appear at AWIC 2011