PHASAR-based Opinion Mining

Master Thesis project plan

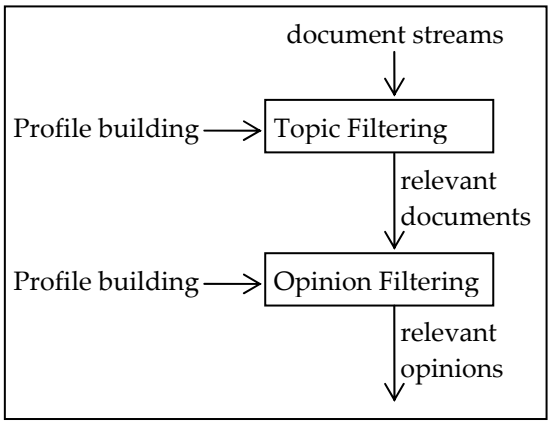
Student:	Jos Claessens (josclaes@sci.ru.nl) s0012823
Supervisors:	Kees Koster, Martijn Oostdijk
External Supervisor:	Raymond Franz (Trendlight B.V.)

October 2006

1 Introduction

This thesis will be about opinion mining, using Phrase-based High Accuracy Search, Analysis and extRaction (PHASAR), developed at the NIII (http://www.phasar.cs.ru.nl). I will briefly introduce this subject now.

An opinion mining system automatically extracts opinions on a given topic out of a set of relevant documents. This set can be manually selected, but more likely it will be the result of a preceding information retrieval system. Although the IR system will probably not achieve 100% precision and recall, thus not providing the perfect set of relevant documents, it is sufficient. At the moment, opinion mining strategies are very basic and the performance is poor. Using more advanced technology, mostly from linguistics, I will try to improve this performance. This will increase the practical usability of specific opinion search in an information retrieval application. The figure on the right shows a typical opinion mining system, consisting of two phases: topic filtering and opinion filtering. PHASAR will be used to achieve the desired performance.



PHASAR is essentially an information retrieval system using linguistic information. Using the AGFL grammar-based language parsing system, it is able to create *dependency triples* out of regular English text. In these triples, pairs of words are stored together with the relation between them. Searching with these triples instead of just words improves results, given a suitable query. Also, due to its interactive nature, users could be able to get closer to what they really search by using suggestions from PHASAR to modify their query. The opinion mining system will be an extension of PHASAR, using the same principles (being based on dependency triples and being interactive).

My research will have an explorative nature. There will be some pilot studies, but the focus is not on building a complete opinion mining application. To carry out pilot studies, I will use data provided by Trendlight, part of TNS/NIPO (http://www.trendlight.nl). These data are media streams on a certain topic and can also be called documents. My work will focus on opinion profile building.

In the remainder of this document, I will specify more exactly what problem I want to solve, how I want to solve it and under what conditions. Also, as far as possible at the moment, I will indicate what literature is available for use.

2 Problem description

This master thesis is about validating or rejecting one "master hypothesis", which is formulated as follows: *'PHASAR-based term extraction is a suitable way to create opinion profiles: it is both efficient and effective'.*

This hypothesis is a very short formulation, in need of some definitions. Most of them are not clear as for now, but defining them will be part of the project.
Of course, what *is* defined, is the first part. PHASAR is a working system (at least for English language documents) and the dependency triple principle has been validated. The PHASAR system will not be modified dramatically. That is: only most required changes may be made, but in principle we will keep it as is.
The definition of *suitable* is given inside the hypothesis: we define term extraction as being suitable when it is both efficient and effective. Still, we will have to deal with those two terms. There is already a clear idea what efficiency and efficacy really mean in this context, though these ideas still need to be concretized. When talking about *efficiency* I will describe in detail how an opinion is isolated within a document. Efficiency is measured using methods commonly used in literature, such as precision and recall and the F_1 value. But first, we will have to work out how the opinion mining system will isolate opinions. In theory, there are various ideas which will have to be worked out and tested in pilot projects. The other part of suitability is *efficacy*. Here it's about determining the optimal browsing and matching strategy. Measures to look at the efficacy of a strategy include time and effort needed to find acceptable results. However, what is considered optimal or acceptable may be subject to change when different users approach the system. Efficacy remains a subjective component of suitability.
A final, but certainly not the last thing to look at is perhaps surprising: what *is* an *opinion* exactly? The system is looking at annotated, but otherwise raw text data. Somehow a notion of what defines an opinion will have to be found. Common sense in practice today is to search for "sentimental" words near words concerning the desired topic in the document. This is not *suitable* enough. But what will be? It will have to be found out.

Besides all these questions, which are mostly on the terrain of information retrieval and linguistics, I will also briefly assess the impact of this technology on the digital society as we know it. When used in certain applications, the advanced search techniques used may lead to various opportunities, in which new problems may arise. In particular, there may be privacy issues involved.

3 Way of working

3.1 Overview

To be able to find answers on my hypothesis, it's necessary to first concretize it. To achieve this, I will study various literature and find out about available definitions of an opinion, matching strategies and browsing strategies. The result will be a concrete definition of what I call a suitable way to make opinion profiles.

Once the hypothesis is fully worked out, the next step is to test PHASAR for its use in this context. In a couple of pilot studies, I will work out how to create opinion profiles for simple opinion search problems. After this phase, it should be clear how PHASAR can achieve suitable opinion profiles. By looking at the theory and validating some promising techniques in pilot studies, I can try to find a good way of working to achieve desired efficiency, while achieving desired efficacy by using optimal matching and browsing strategies.

Finally, the system should be able to withstand the "real life" test, processing the processed streams from Trendlight and producing valuable opinion profiles and relevant opinion data.

After literature study and practical (pilot) study is completed, it's time to reflect on the achievements. I will look at the new possibilities and their effects. Also, I will look at what is not (yet) possible or what looks promising for the future.

3.2 Relevant resources

The literature to be used will be found mainly in the area of information retrieval and linguistics. Especially when dealing with the question what an opinion is, linguistics will prove useful. Both areas will contribute in defining and achieving high efficiency and efficacy.

Already working systems like AGFL and PHASAR will be used, together with their associated literature and developers' knowledge. If needed, there will be contact with the linguistics department of the university.

Finally, Trendlight will provide relevant test data streams and knowledge.

3.3 Schedule

The project should be completed in six months. Start date is the first of October 2006, finishing date is the first of April 2006. I strive to keep to the following schedule:

October 2006	Literature study; starting up.
November 2006	Pilot study: create a profile for a simple opinion mining problem. The process: query → generalized query → profile for 1 opinion.

December 2006	Find answers to research questions, concretizing the central hypothesis. Defining efficacy and efficiency measures, elaborating the "way of working".
January 2006	Application of pilot system to Trendlight material, processed by Topic Filtering system by Johan Stortelder.
February 2006	Extension period; start finalizing thesis.
March 2006	Finalize thesis.

This schedule is still quite rough and will be updated during the project. At this moment, we planned a relatively long finalizing period, to be able to shift planning deadlines according to advancing knowledge of realistic progression.

3.4 Contact

To keep in touch with others working on this project and its surroundings, I will have a meeting with my main supervisor, Kees Koster, on a weekly basis. Furthermore, I will contact my other supervisors as needed during the project.