University of Cambridge

Work plan:
Personalized Search

Nicolaas Matthijs MPhil Computer Speech, Text and Internet Technology Supervised by Dr. Filip Radlinski (Microsoft Research) and Dr. Stephen Clark

Personalized Search

1. Introduction

Different people look for different things with the same search queries. Classic examples include queries like jaguar, Columbia, flash, mercury, ... This makes it harder for search engines to come up with results that are relevant to a user's current information need. My research project will attempt to learn a user's interests by looking at different sources of personal data. The final output is a personalized search system that will aim to bring the results of a search engine relevant to that user to the top of the search ranking, which should make searching faster and allow for more relevant discovery .

2. Aims of the project

Core goals:

- Development of methods and technologies that make it possible to capture data sets and evaluate results in personalized search projects.
- Development of 1 or more algorithms that are able to extract keywords (and more specifically interests) from data related to the user in HTML format. The project should also assess which algorithm is the most successful and effective in this task.
- Development of a method that can re-rank the output of an online search engine given information about a user's interests.
- Development and dissemination of a personalized search browser add-on that makes it possible to use the developed technologies and algorithms on a daily basis.

Additional goals:

- Assessment of which personal data source (browser history, e-mail, files on HD, FaceBook profile, ...) describe a user and his interests best.
- Produce a publication that talks about the ideas, approaches and outcomes of this project.

3. Implementation/Approach

Because there are no publicly and readily available corpora and data sets for personalized search, the first phase of the project will involve capturing sufficient data for creating such a data set. As the project will initially look at using browser history for personalization, I will write a Firefox add-on that will capture this. I'll then go out to fellow students and ex-colleagues in order to recruit them for using the add-on to allow me to capture sufficient data. The add-on will also try to keep privacy into account by not capturing any data that requires authentication.

In the second phase of the project I will do some background reading through a selection of papers in personalized search, keyword extraction, information extraction and topic classification. I'll also do some research into existing personalized search products (like SurfCanyon) and I'll evaluate and compare the methods they're relying on and the technologies they use to deliver their product to real users.

The third phase will contain most of the actual development. First of all, I'll normalize the database to take out login pages and extract the text from the collected HTML. Most of the time will then be spent on developing 1 or more algorithms to extract keywords and user interests from the given data. In the first instance, these interests will be used to predict which Wikipedia disambiguation is most useful to a user. Next, I will attempt to re-rank the output of a search engine to make the ranking more biased towards the user's interests. If time permits, I'd also like to collect different personal data sources like e-mail, files on HD, FaceBook profile, ... and run the same interest extraction technology on those.

In the final phase of the project I will evaluate how well the algorithm does on the held-out test data. If I managed to use different data sources, I'll also assess which one describes the user's interests best. Next, I will develop a browser add-on that will actively re-rank the output of search engines and can be used on a daily basis by real users. I'll also build a website for the add-on, so that it can easily be disseminated. If time permits, I would also like to do some life user testing of the developed technology at the Computer Labs.

4. Special requirements and dependencies

Hardware:

I will use my personal laptop for all of the development. I have got about 100Gb of disk space available for this project. Specifications: MacBook Pro 2.66 GHz, 4 Gb RAM, 250 Gb Disk, Mac OSX Snow Leopard.

Back-up:

I will take care of back-ups by using an external hard drive and Time Machine.

Web server:

A regularly backed-up Apache web server and public domain (http://alterego.caret.cam.ac.uk) have been made available by the Center for Applied Research in Educational Technologies at the University of Cambridge.

Data storage:

A MySQL database has been made available by the Center for Applied Research in Educational Technologies at the University of Cambridge. This will hold all of the test and training data.

No special resources from the Computer Labs are required.

5. Time Schedule

Data capturing

- Add-on development (16th - 21st of February): Development of a Firefox add-on that will capture people's browsing history. This will provide me with training and test data for the project.
- Website development (20th - 23rd of February): Development and rollout of a website that allows people to download the Firefox add-on and gives them more background information about the project. People will also be able to refer their friends and/or colleagues to this website.
- Recruitment (22nd - 25th of February): Recruitment of people that will install the Firefox add-on. These will mainly be fellow students and ex-colleagues.
- Data capturing (23rd February - 1st of April): The time period over which browsing data will be actively captured by the Firefox add-on. After that, the add-on will stop capturing data.

Background reading

- Personalized search papers (17th - 19th of March): Reading a selected set of papers about personalized search.
- Keyword extraction papers (19th - 24th of March): Reading a selected set of papers about keyword extraction and keyword extraction from web pages.
- Information extraction papers (24th - 27th of March): Reading a selected set of papers about information extraction and topic classification.
- Analysis of similar products (29th - 31st of March): Searching for and reading about products and technologies that do something similar to the goal of this project and are already available and usable (SurfCanyon, ...).

Main implementation

- Database normalization (1st - 4th of April): Removal of advertisements and login pages. Extracting the text from the HTML source codes and removing the HTML tags.
- Extraction of the training and development set (4th - 6th of April): Dividing the collected data into training set, development set and test set. Finding a way in which quick evaluation can be done during development.
- User interests extraction algorithm (6th - 26th of April): Development of 1 or more algorithms that can be used to extract a user's personal interests by looking at a stream of data specific to that user.
- Application to Wikipedia disambiguation (26th - 29th of April): Predict which Wikipedia disambiguation is the most interesting to a user given his interests.
- Application to search engines (29th - 6th of May): Using the interests learned by the extraction algorithm to re-rank the output of a search engine.
- Capturing of alternative data source (6th - 11th of May - optional): Try to get hold of alternative personal data sources like e-mail, files on HD, FaceBook profile, ... from people that have run the add-on as well.

Evaluation

- Development of the test set (11th - 13th of May): Setting up scrips to run and evaluate the developed algorithms on the test set. This will involve preparing the test set to make this process easier.
- Running technology on training and test set (13th - 15th of May): Evaluating how well the algorithm does on both the complete training and test set.
- Evaluation on alternative data sources (17th - 22nd of May - optional): Evaluating how well the same algorithm does on different types of personal data.
- Creation of browser add-on (22nd - 30th of May): Development and dissemination of a browser add-on that captures your personal data, learns what the user's personal interests are and re-ranks the output of several search engines to be more biased towards the user's personal needs.
- User testing (30th of May - 2nd of June - optional): Bringing people into the computer labs and have them test the developed browser add-on in real life.
- Performance optimization (30th of May - 2nd of June - optional): If there is an issue with the developed add-on in terms of performance, I'll try to optimize it so it can be used by people on a daily basis.

Writing

- Finish writing up Masters thesis (1st - 16th of June): Writing down the results from the evaluation and finish writing up the details about the developed algorithm.
- Creating publication (10th - 17th of June - optional): If time permits and the output of the project is found to be useful, a publication can be created from the Masters thesis.
- Handing in thesis (17th of June)

Project presentation:

- Presentation creation (21st - 28th of June): Create a 15 minute presentation that summarizes the goals, approaches, algorithms and evaluation results of the project.
- Presentation (30th of June): Project presentation at the Computer Labs

External Supervisor:	Internal Supervisor:
______________________________	______________________________
Dr. Filip Radlinski	Dr. Stephen Clark