ERCIM "Alain Bensoussan" Fellowship Scientific Report

Fellow: Eugen Popovici
Visited Location : CWI, Amsterdam, The Netherlands
Duration of Visit: 01/03/2009 – 30/11/2009

I - Scientific activity

During my first nine months visiting period as an ERCIM fellow within the Interactive Information Access Group at CWI in Amsterdam I have explored the following research fields: Named Entity Recognition, Semantic Web Technologies, User Studies, News/eBook Navigation and Image-Text Relationships. During this period I have initiated and worked on two research projects:

1. *eBook Library Explorer*: aiming to make use of the 'recognized' named entities linked to semantic data and 'appropriate' visualization techniques to explore eBook collections.
2. *enrichme: ENRICHhing docuMEnts with annotated images* - aiming to place annotated images within text documents.

eBook Library Explorer

Most of the current portals for eBooks provide simple interfaces and interactions based on keyword search, access to the table of contents and to the document metadata.

Users explore (browse) the content of a collection/book in contexts where reading the whole book may not be appropriate: I) during the book selection process, II) when the user is interested in getting an overview of the content of a part of a book, III) find, cite, quote or IV) compare facts (about the characters, geographical locations, the time periods) at different (sub)collection/book structural levels.

The aim of this project is to provide a visual framework/prototype for exploring ('more deeply') the content of eBook collections. It proposes to visualize named entities (characters, geographical locations, dates) within their textual and structural context and to explore their semantic relationships. We hope that this will stimulate curiosity and encourage exploration so that the user could make opportune discoveries. It also aims to allow access to eBooks content in a non-sequential manner and to support comparative analysis between books or books fragments focusing on the recognized named entities and their 'context'.

An initial prototype was developed based on the *Open Calais Web Service*[1] for recognizing named entities and on the *NewsML and the Semantic Web*[2] project architecture for disambiguating and linking the recognized named entities to the *DBpedia*[3] and the *GeoNames*[4] geographical database. The current implementation of the graphical user interface allows visualizing and comparing geographical locations at different book structural levels based on the *Google Maps APIs*[5]. The next phase of the project involves adding adequate visualizations for each potentially interesting named entity type and finding an appropriate group of users and use cases for evaluating the system.

[1] http://www.opencalais.com/
[2] http://newsml.cwi.nl/
[3] http://dbpedia.org/
[4] http://www.geonames.org/
[5] http://code.google.com/apis/maps/

enrichme: ENRICHing docuMEnts with annotated images[6]

Enriching text documents with images is a common, time consuming, and cognitive intensive task in the day to day activities such as writing a school report or publishing a new entry on a Blog. Automatic methods that assist the user to find relevant images for the document content and place them within the document have the potential to reduce the user's time spent on searching, selecting and organizing the images within the document.

This project uses and evaluates text-based matching techniques and results from the Focused Information Retrieval research field (XML IR and Passage IR) to find the appropriate locations for placing annotated images within a text document. The proposed locations should take into account both the proximity of the images to the relevant text fragments and the global organization of the images at the document level.

To automatically enrich a document with images, we propose to compute the relevance of an annotated image to a document fragment based on a language modeling approach. To select and place the images within the document we construct a bipartite graph having as vertices images and document fragments linked by edges showing the relevance between each image and each document fragment. Common graph centrality measures such as PageRank are used to select images that cover well the document content and relate to the most relevant document fragments.

We are currently conducting an initial user study on how users enrich documents with annotated images. This will also allow the creation of a golden standard for evaluating image placing strategies. Based on the analysis of the data gathered during the initial user study, a user-model will be derived and incorporated in the automatic method for selecting and placing the images within the documents.

The next phase of the project is to evaluate the effectiveness of the automatic image placing strategy against the golden standard completed with a qualitative evaluation by the users. The future research plans for this project include an evaluation of the adequacy of the "document enriched with images" metaphor to discover and visualize the relations between documents and images and serendipitously explore the content of multiple text and multimedia collections. Another research direction is the development of recommendation algorithms for finding i) related images relevant for the document context and ii) relevant documents for a subset of selected images.

II- Publication(s) during your fellowship

Two CWI technical reports and a conference paper are under preparation:

1. "*Building an evaluation framework for image placement strategies within text documents"*, CWI Technical Report.
2. "*Investigating users' behavior while enriching documents with annotated images"*, CWI Technical Report.
3. *"A Graph-Based Strategy for Enriching Documents with Annotated Images*", Conference paper.

Other professional activities

1. Reviewer for the Information Processing & Management - IP&M International Journal, Elsevier.
2. Member of the posters program committee for the 32nd Annual ACM SIGIR Conference, Boston, July 19-23 2009.

[6] http://media.cwi.nl/enrichme/about.html

III -Attended Seminars, Workshops, and Conferences

1. *"Dutch Semantic Web Get-together"* Workshop, Vrije Universiteit, Amsterdam, Netherlands, March 16th, 2009.
2. *"The Unreasonable Effectiveness of Data"* workshop and panel discussion, Delft, Netherlands, Friday, May 8, 2009.
3. *"Leesclub"* seminar series of the Interactive Information Access Group, CWI, Amsterdam, Netherlands, March-November 2009.
4. *"Data Explosion"* seminar series, CWI, Amsterdam, Netherlands, March-November 2009.

Attended Courses and Tutorials

1. *"Probabilistic Methods for Entity Resolution and Entity Ranking"*, Advanced course of the Netherlands research School for Information and Knowledge Systems - SIKS, Conference Center Woudschoten, Zeist, Netherlands, April 20th– 21th, 2009.
2. *"A Semantic Multimedia Web: Create, Annotate, Present and Share your Media"* tutorial by Lynda Hardman (CWI & UvA) and Raphael Troncy (CWI), University of Amsterdam, Netherlands, May 15th, 2009.