

# Student Exchange Program

## University of Western Australia, School of Computer Science

### Diploma Theses Proposals 2010 Semester 1

#### Text Analysis Tool

The proliferation of interactive Web portals such as online forums, message boards and blogs, has opened a new era of human communication. The incredible accessibility and anonymity has made the Web a fertile ground for opinionated and directional text. Consequently, the power of the “word of mouth” and the “voice of the people” have accelerated to an unprecedented new level and scale. This is both a blessing and a curse to the governments and the businesses. On the one hand, it has never been this easy to gather public opinions, be it on goods, services, brands, reputations, or their political standings on politicians, policies and hot issues such as Climate Change. On the other hand, the amount of information available is incomprehensible. For example, it is easy to come across 300+ reviews for just *one* single product within *one* portal, a SONY camera on Amazon. The reviews can be lengthy, sometimes thousands of words. Monitoring and making sense of this enormous amount of data at an executive level is an extremely challenging task. Yet failures in moderating a forum to leave rooms for rumours from some powerful online “no-one” can lead to disastrous situations. The recent arrest of a 30-year-old unemployed South Korean<sup>1</sup> for spreading pessimistic financial forecasts on his blogs indeed signifies an urgent need in effective monitoring of online public opinions.

In this project, we propose to use information science, in particular, **text mining**, to help decision makers digest, monitor and even make effective use of these myriad online voices so as to maximise the socio-economic benefits.

**The research project focuses on Text Cube, more specifically along the lines of building RDF data warehouse for Online Analytical Processing for Text.**

It is suggested that data warehouse design should go through conceptual, logical and then physical stage [1, 2]. Following a relational model of data warehousing, in this project at the *conceptual level*, we will first identify the dimensions, which can be represented as dimension tables, and then link the dimensions through a fact table. Such a multi-dimensional data model can be visualised as data cubes. Traditionally, a document is treated as a “bag of words”. For the *logical level*, we propose a novel view of a document. We consider a document as decomposable and can be represented by a data cube along three dimensions (social, semantic and sentiment). Each point of a data cube (i.e. document) in the text warehouse is an event. Each event is a *Social entity's Sentiment orientation* towards several *Semantic dimensions* of an object, at certain point of *time* and with specific *geographical* references. Both time and geo-reference will be classified either as *source* or *target* reference. At the *physical level*, this research chooses a distributed RDF data warehouse due to the fact that some relations, such as the social and semantic relations, are best expressed as linked graphs. We will review the available RDF data warehousing packages (e.g. Virtuoso Open Source) and select from the ones that are popularly adopted and verified in the bioinformatics domain to deploy our storage facilities.

---

<sup>1</sup> <http://www.guardian.co.uk/world/2009/jan/22/south-korean-blogger-minerva-prosecution>

Further information can be found at [http://www.cs.uiuc.edu/homes/hanj/pdf/icdm08\\_xlin.pdf](http://www.cs.uiuc.edu/homes/hanj/pdf/icdm08_xlin.pdf) and [http://en.wikipedia.org/wiki/Online\\_analytical\\_processing](http://en.wikipedia.org/wiki/Online_analytical_processing)

- [1]. Prat, N., Akoka, J., & Comyn-Wattiau, I. (2006). A UML-based data warehouse design method. *Decision Support System*, 42 (3), 1449--1473.
- [2]. Rizzi, S., Abello, A., Lechtenborger, J., & Trujillo, J. (2006). Research in data warehouse modeling and design: dead or alive? *DOLAP '06: Proceedings of the 9th ACM international workshop on Data warehousing and OLAP* (pp. 3--10). New York, NY, USA: ACM.

Supervisor UWA: Wei Liu

Supervisor TUG: Christian Gütl

### **Procedure for Pre-selection (must be written in English)**

1. Prepare a Short CV including interests, programming and software design skills, practical work experiences, language skills and other relevant information
2. Progress Report (Overview) of the study program and average performance level
3. Short Application Letter (1 page) stating motivation/reasons why to apply for the Thesis position and outlining interest in the topics of the theses proposals

### **Procedure for Scholarship**

1. Letter of Acceptance from supervisor of Host University (result of preselection)
2. Application for Scholarship at Graz University of Technology

### **Time Schedule**

- October 20<sup>th</sup>, 2009 Application as PDF File via email to [cguetl@iicm.edu](mailto:cguetl@iicm.edu)
- October 21<sup>th</sup>, 2009 Preliminary decision by supervisors and letter of agreement for supervision by CBS, School of Information Systems, Curtin University of Technology
- October 31<sup>st</sup>, 2009 Application for scholarship  
Claudia Buchrieser, Office of International Relations, Graz University of Technology
- Master Thesis at Host University  
from Februar 2010 until July 2010

### **Further Information:**

#### ***Local Contact:***

Christian Gütl, [cguetl@iicm.edu](mailto:cguetl@iicm.edu)

#### ***Information about scholarship:***

Barbara Recla, [barbara.recla@TUGraz.at](mailto:barbara.recla@TUGraz.at)

[http://portal.tugraz.at/portal/page?\\_pageid=133.1&\\_dad=portal&\\_schema=PORTAL](http://portal.tugraz.at/portal/page?_pageid=133.1&_dad=portal&_schema=PORTAL)

#### ***Information about University of Western Australia***

<http://www.uwa.edu.au/>