CS - 4200 Programming Project

Project Proposal

Group Members:	Jayasinghe D.S.	060199T
	Abeywickrama G.P.S.P.	060012R
	Ketteepearachchi D.C.	060235D
	Hettiarachchi S.	060168A

Department **:** Computer Science & Engineering

Project Title **:** Text clustering based concept hierarchy to generalize from different text sources

Supervisors **:** Dr Damminda Alahakoon (external)
Mr. Sumith Matarage (external)
Miss. Upuli Gunasinghe (internal)

Introduction

Living in a modern and technologically dependent world, we heavily rely on electronically stored data and information to come up with sound and timely decisions. In this context, data mining and retrieval processes hold a critical responsibility. Therefore even a slight improvement in their effectiveness could cause a revolution in the IT environment.

In any establishment, its success essentially depends on the decision making process followed. Reliable and accessible data would be the driving force of the aforementioned process. In the current scenario, although the data is available in an unimaginable volume, the scattered nature of relevant and non-relevant data hinders the data retrieval process.

Our aim is to come up with a sound and concrete solution so that the gigantic volume of unsorted data could be fitted in to a virtual hierarchy to make extracting relevant data more feasible. A data hierarchy will enable the data requesting party to track down specific data clusters directly saving an enormous amount of valuable time and effort.

Problem Definition

In the past several years, information technology created lots of innovations in various fields such as business, education, health, defense, etc. Everyday organizations from these areas collect high quality data on a large scale from many different sources. The huge amount of data can be a gold mine for organizational management. It is therefore increasingly important to build a central information processing unit that merges all the different sources and thus provides us a greater advantage in information processing and decision making. However, timely and accurately processing tremendous data analysis in traditional methods is a difficult task. The ability to analyze and utilize massive data lags far behind the capability of gathering and storing it. This gives rise to new challenges for businesses and researchers in the extraction of useful information.

Consider a very large collection of textual items, such as an encyclopedia or a digital library where data is accumulated from different sources. It would be of great help for browsing it, if the items could be pre-ordered hierarchically according to their contents. For the ordering one needs a similarity measure for the pairs of items. One might wish to have a measure that compares the meanings of the contents linguistically. When the text corpus is really large, such linguistic analyses become soon computationally overwhelming. It has transpired, however, that rather descriptive and useful similarity relations between text items are already reflected in the use of the words in them.

Therefore our aim is to develop a user-friendly and an efficient method and a tool that will satisfy our different kinds of information needs and tasks regarding organizing, visualizing, searching, categorizing and filtering textual data.

Proposed Solution

A huge advantage could be attained in data mining and information retrieval processes by placing the vast volume of available unsorted raw data in a meaningful hierarchy. To obtain this functionality the technique of text clustering could be used. Using the method of text clustering a given set of text documents could be divided in to a number of sub sets based on their contextual similarity.

First step of text clustering is to assign a vector for each document to create a way to compare and identify similarities of different documents. To obtain the document vector, first the document is filtered to remove stop words such as articles, conjunctions, prepositions, etc. which bear no content information. Use of thesauri could aid in defining synonyms to enhance the content information gathering. Then based on a statistical analysis of the input document set, a set of key words or index words are chosen to create document vectors. Document vector for each document is generated by calculating the number of occurrences of index words in each document [1]. Then these document vectors are used in an unsupervised clustering algorithm such as Self Organizing Maps [2], [3], Growing Self Organizing Maps[4], k-means, etc. to compare documents and come up with a set of document clusters which contains similar documents [1]. A literature survey has to be carried out to decide on the optimum algorithm which could be used to obtain the document clusters.

By applying above steps again for a cluster of documents, a refined set of clusters could be obtained from the original cluster of documents. By applying this procedure repeatedly, a meaningful hierarchy could be obtained from the initial set of unsorted text documents.

Scope of the Project

We find various kinds of information sources, which remain in an unsorted manner at any modern organization. Informative data, which we can use for clustering process comes in many different forms like web documents, text documents, images, videos, etc. However, here the focus of this project is to implement a system, which can cluster a set of text documents. In further details, we use a set of text documents with specified maximum length as the input to the system and after the clustering process; the system should be able to categorize those documents in a hierarchical manner. Actually, this should be a tree-based hierarchy with several layers and the final output would be a folder hierarchy, which represent above mentioned tree based hierarchy. We leave the number of layers in the hierarchy for the user to decide and it would be a parameter of the system. At the end of the day this whole process would facilitates a user to find a particular document he/she is interested in from a large pool of text documents.

For the clustering purpose, we might use a part of the text document since it would consume a huge amount of computational power and memory if we use the whole document. Therefore we may take the introduction or the table of content of a text document and abstract or the introduction of a research paper to build up the input. Here we give the flexibility to the user by making this fact also a parameter of the system.

Time-line of the Project

ID	Task Name	Start	Duration	Q3 09			Q4 09			Q1 10			Q2 10		
				Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
1	Literature Survey	7/1/2009	12w	■	■	■									
2	Design & Analysis	8/3/2009	6w		■	■									
3	Implementation of Text Extraction interface from Input Source	9/14/2009	2w			■									
4	Implementation of Data Pre-processor	10/1/2009	5w				■	■							
5	Implementation of the clustering algorithm	11/2/2009	8w					■	■						
6	Implemention of other extended features	1/1/2010	6w							■	■				
7	Testing and Integration	2/16/2010	2w								■				
8	Writing Documentation and Research Papers	4/1/2010	6w										■	■	

References

[1] A. Klose, A. N¨urnberger, R. Krus and G. Hartmann, M. Richards, Interactive Text Retrieval Based on Document Similarities, *Phys. Chem. Earth*, Vol. 25, No. 8, pp. 649–654, 2000

[2] Kohonen, T., Hynninen, J., Kangas, J., and Laaksonen, J. (1996a). SOM_PAK: The self-organizing map program package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo.

[3] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59-69.

[4] Alahakoon, D., Halgamuge, S., and Srinivasan, B., A structure adapting feature map for optimal cluster representation, in *Proc. Int. Conf. On Neural Information Processing*, pp. 809–812, Kitakyushu, Japan, 1998.