UNIVERSITI PUTRA MALAYSIA

INCORPORATION OF CONTEXTUAL RETRIEVAL AND DATA FUSION APPROACH TOWARDS IMPROVING THE RETRIEVAL PRECISION

AZ AZRINUDIN ALIDIN.

FSKTM 2007 18

INCORPORATION OF CONTEXTUAL RETRIEVAL AND DATA FUSION APPROACH TOWARDS IMPROVING THE RETRIEVAL PRECISION

By

AZ AZRINUDIN ALIDIN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

November 2007

DEDICATION

I want to dedicate this thesis to my loving family and to 'my special one'.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

INCORPORATION OF CONTEXTUAL RETRIEVAL AND DATA FUSION APPROACH TOWARDS IMPROVING THE RETRIEVAL PRECISION

By

AZ AZRINUDIN ALIDIN

November 2007

Chairman: **Shyamala Doraisamy, PhD**

Faculty: **Computer Science and Information Technology**

Generally, the functionality of information retrieval (IR) could be divided into two categories where one section deals with search and retrieval while the other component concerns with the subject or content analysis. In the search and retrieval part, the IR systems present a ranked list of relevant documents depending on the user submitted query as the representation of the user's information need. The ranked list given indicates the probability of the document is relevant to the query by ordering the highest relevant document at the top position and so forth. However, queries are often formulated with simplified short words, such as "Java". These words are unable to summarise precisely the user's information need and its context, i.e. "java, programming language" or "java, the island". Consequently, the user's information need is not satisfied as the highest relevant document was not positioned accordingly or too much relevant document was presented in the ranked list.

Besides, by using the simplified query made the context is not easily extractable, and in recent years there has been much research interest in contextual retrieval. Likewise

IR, contextual retrieval retrieved the relevant document by using the combination of query, user context and search technology into a single framework. Furthermore, in contextual retrieval, the user's context is exploited to differentiate the relevant document that is useful at that time the requests occur.

On the other hand, in order to match the queries and the document representation, different IR schemes were applied to calculate the probability. As a result, often retrieval precision is different for differing IR schemes, where dissimilar lists of relevant documents for the same query submitted are presented. Thus, data fusion approach is implemented in the IR to overcome this complication where multiple sources of results are combined. The implementation of data fusion approach in IR involves the merging of retrieval result from different IR schemes into a single unified ranked list that supposedly presents a list of high precisely relevant document.

This study presents an approach to incorporate contextual retrieval and data fusion by using a one-keyword query towards improving retrieval precision. The methods to identify user context are categorised into four approaches; relevance feedback, user profiles, word-sense disambiguation and knowledge engineering. In order to extract user context and to model contextual retrieval, term-weighting scheme based on user profiles and knowledge engineering approaches for Watson scheme and word-sense disambiguation approach for WordSieve scheme are implemented in this study. Five randomly selected documents are selected and submitted to these schemes and the user's context extracted is used to expand the initial query for retrieval process.

In addition, the feasibility of adopting a data fusion approach was assessed in this study by testing two preconditions; —the efficacy and dissimilarity tests for the IR scheme candidates, as there is a possibility that the precision improvement may not be accomplished. Two queries which are Java and Jaguar, expanded by using user's context extracted by Watson and WordSieve are submitted and more than ten thousand documents are collected as the data collection for conducting the experiment. The performance of the experiment is evaluated by using three assessments; precision recall graph, precision evaluation based on document ranked and mean average precision. The data fusion experiment based on contextual retrieval results has reveals significant improvement on retrieval precision where the lowest percentage gained compared to the basic IR scheme is approximate to thirty seven percent, ten percent improvement compared to Watson and fifthteen percent improvement compared to WordSieve based on mean average precision calculation.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PENGGABUNGAN DAPATAN SEMULA BERPANDUKAN KONTEKS DAN PENDEKATAN PELAKURAN DATA UNTUK MENINGKATKAN KETEPATAN DAPATAN SEMULA MAKLUMAT

Oleh

AZ AZRINUDIN ALIDIN

November 2007

Pengerusi: **Shyamala Doraisamy, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Secara amnya, fungsi sistem dapatan semula maklumat boleh dibahagikan kepada dua kategori di mana satu komponen berfungsi untuk melakukan proses mencari dan mendapatkan semula maklumat manakala satu komponen lagi melakukan analisis terhadap subjek atau kandungan dokumen. Dalam komponen mencari dan mendapatkan semula maklumat, sistem dapatan semula maklumat akan menyenaraikan kedudukan dokumen yang relevan bergantung kepada pertanyaan yang dihantar oleh pengguna sebagai pengganti kepada kemahuan maklumat pengguna. Senarai kedudukan yang diberikan menunjukan kebarangkalian bagi sesuatu dokumen itu relevan kepada pertanyaan pengguna dengan meletakkan dokumen yang menpunyai darjah relevan tertinggi di kedudukan teratas dan di ikuti oleh kedudukan seterusnya. Walau bagaimanapun, penggunaan perkataan yang ringkas dan pendek seperti "Java" selalu digunakan dalam pembentukan pertanyaan. Penggunaan perkataan seperti ini menyebabkan kemahuan maklumat pengguna dan juga konteks perkataan tidak dapat difahami, i.e. "java, bahasa pengaturcaraan" atau

"java, kepulauan". Disebabkan itu, kemahuan maklumat pengguna tidak dapat dipenuhi kerana dokumen yang relevan tidak diletakkan di kedudukan yang sepatutnya dan terlalu banyak dokumen yang relevan tersenarai.

Disebabkan oleh penggunaan perkataan yang ringkas di dalam pertanyaan pengguna tidak dapat menunjukan konteks perkataan, beberapa tahun kebelakangan ini kajian berkenaan dapatan semula berpandukan konteks telah mendapat perhatian. Seperti juga dapatan semula maklumat, dapatan semula berpandukan konteks mendapatkan dokumen yang relevan dengan menggabungkan pertanyaan pengguna, konteks pengguna dan teknologi carian dalam satu rangka kerja. Tambahan pula di dalam dapatan semula berpandukan konteks, konteks pengguna diexploitasi untuk membezakan dokumen yang sebenarnya relevan ketika masa permintaan itu berlaku.

Namun begitu, untuk memadankan pertanyaan pengguna dan perwakilan dokumen, terdapat pelbagai skima dapatan semula maklumat yang boleh digunakan untuk menghitung kebarangkalian tersebut. Disebabkan itu, hasil ketepatan dapatan semula maklumat selalu berbeza untuk skima dapatan semula maklumat yang pelbagai di mana senarai dokumen yang relevan yang tidak sama dihasilkan bagi pertanyaan pengguna yang sama. Maka dengan itu, pendekatan pelakuran data telah pun digunakan dalam dapatan semula maklumat untuk mengatasi komplikasi ini di mana kepelbagaian sumber keputusan digabungkan. Pendekatan pelakuran data ini diimplementasikan di dalam dapatan semula maklumat dengan melibatkan cantuman keputusan yang berbeza dari skima dapatan semula maklumat untuk menghasilkan senarai penyatuan tunggal yang mempunyai relevan dokumen yang berketepatan tinggi.

Kajian ini mempersembahkan pendekatan untuk menggabungkan dapatan semula berpandukan konteks pengguna dan pelakuran data dengan menggunakan satu perkataan dalam pertanyaan pengguna untuk meningkatkan ketepatan keputusan dapatan semula maklumat. Kaedah-kaedah untuk mengenalpasti konteks pengguna telah dikategorikan kepada empat pendekatan; suap-balik relevan, profil pengguna, perkataan pelbagai makna dan kejuruteraan pengetahuan. Untuk memodelkan dapatan semula maklumat berpandukan konteks, skima pemberat perkataan berdasarkan pendekatan profil penguna dan kejuruteraan pengetahuan untuk Watson dan pendekatan perkataan pelbagai makna untuk WordSieve telah pun diimplementasikan. Sebanyak lima dokumen yang dipilih secara rawak telah digunakan ke atas skema-skema ini dan konteks pengguna yang diperolehi digunakan untuk mengembangkan pertanyaan asal untuk melakukan proses dapatan semula maklumat.

Kajian ini juga telah menilai kebolehjadian untuk mengadaptasikan pendekatan pelakuran data dengan menguji dua pra-syarat; –ujian keberkesanan dan ujian perbezaan bagi calon-calon skima dapatan semula maklumat kerana terdapat kemungkinan peningkatan ketepatan tidak akan diperolehi. Dua jenis pertanyaan pengguna iaitu *Java* dan *Jaguar*, dan dikembangkan oleh konteks pengguna yang diperolehi daripada skema Watson dan skema WordSieve telah digunakan dan lebih daripada sepuluh ribu dokumen telah dikumpulkan sebagai koleksi data untuk mengendalikan eksperimen ini. Prestasi eksperimen ini dinilai dengan menggunakan tiga jenis penilaian; graf ketepatan perolehan kembali, penilaian ketepatan berdasarkan susunan dokumen dan min purata ketepatan. Eksperimen pelakuran data dengan menggunakan keputusan dapatan semula maklumat berpandukan konteks

mendedahkan peningkatan yang berkesan bagi ketepatan dapatan semula maklumat di mana nilai peratusan terendah yang diperolehi dengan skima dapatan semula maklumat asas sebagai perbandingan ialah menghampiri tiga puluh tujuh peratus, dengan sepuluh peratus peningkatan bagi Watson dan lima belas peratus peningkatan bagi WordSieve bagi mengiraan berdasarkan perbandingan min purata ketepatan.

ACKNOWLEDGEMENTS

Special thanks to:-

My supervisor, Dr. Shyamala Doraisamy,

My co-supervisor, Assoc. Prof. Dr. Hjh. Fatimah Dato Ahmad,

My colleagues at FSKTM,

My family,

last but not least,

'My Special One'

Thanks again for being so helpful, supportive, understanding and have faith in me to complete this Master thesis

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF TERMS

The terms used in this thesis may carry a different interpretation. The definition of terms use in this whole thesis is listed below:

Collection	Set of documents
Document	Unit of indexed text (documents, paragraphs, dictionary entries, web pages)
Information retrieval (IR) scheme	The information retrieval (IR) technique that differs in its parameter setting.
Observation	The outcomes of the experiment which could be measured.
Query	The user's information need, expressed as a set of terms.
Ranked List	A list of relevant document retrieved which is ranked based on its relevance to the query.
Similarity Computation	A matching process between query and the document by using various IR techniques.
Term	A lexical item or phrase that occurs in collection.
Term Weight	A weighted value for term in document and the whole collection.
Treatment	A modification that intent to enhance the IR process.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

> "An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his enquiry; it merely informs him of the existence (or non-existence) and whereabouts of documents relating to his request" – Lancaster (1968)

1.1 Research Background

Information retrieval (IR) deals with representation, storage, organisation of, and access to information items. Information stored in IR should be easily accessible if user requires for that information. In early days of IR, only text information was organised and stored but it was a major advance in libraries in assisting user to locate and selecting the right information. Today, the scope of IR function not only serves users in libraries environment but also world communities since the World Wide Web and Internet are available to everyone. Information stored also changes from textual information only to multimedia information consists of text, video, images and audio. Despite the transformation of information stored yet the functionality of IR still remain the same, to retrieve relevant document or information required by user community.

The overview of the IR purpose discusses above indicate that one side of IR system stored information representation and on the other side deals with user queries. Presenting a ranked list of relevant information is a process that linked up these two sides as depicted in Figure 1.1

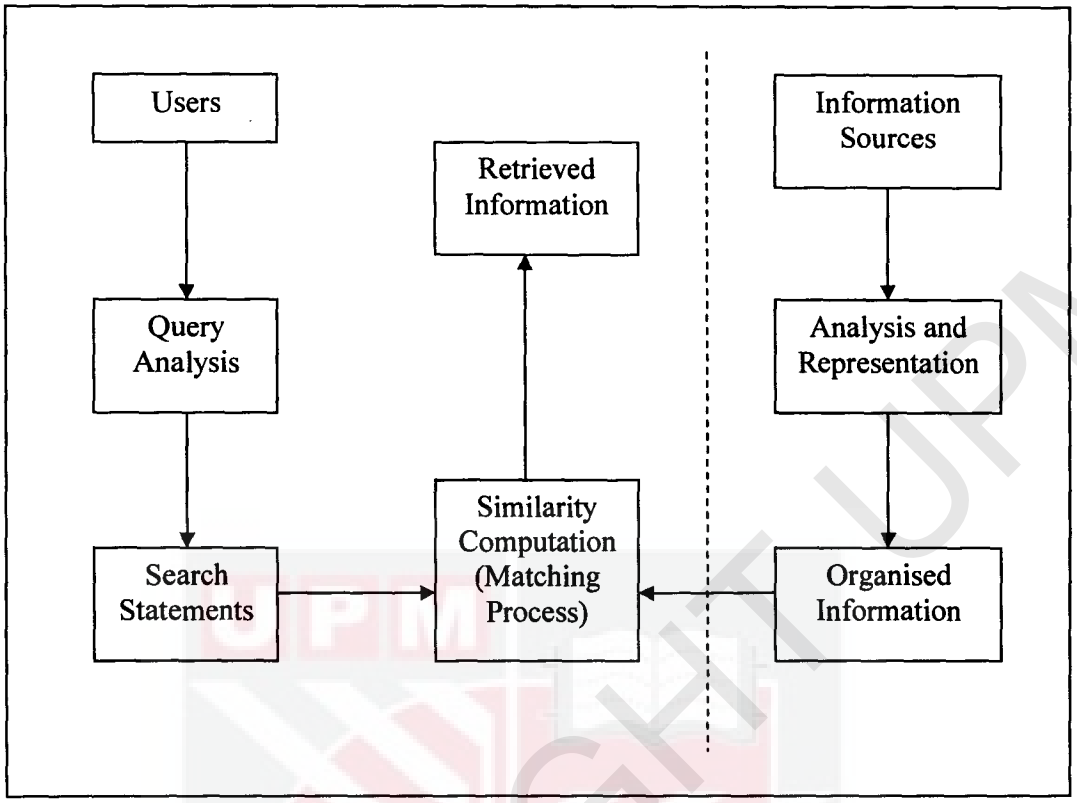

Figure 1.1 Information Retrieval System Components by Chowdhury

Figure 1.1 shows that all tasks in IR can be separated into two major groups – subject/content analysis, and search and retrieval. The tasks correlated to the analysis, organisation and storage of information are includes into subject/content analysis. The search and retrieval process involves the tasks of formulating search statements from analysing user's queries, the matching process for searching, retrieving and producing the ranked list of relevant information. Research in IR also falls into these groups where the IR performance can be improved by designing methods to identify, analyse and organise information on subject/content analysis group and develop searching techniques on search and retrieval group.

Generally, IR system is designed to retrieve the relevant information requested by user. In order to retrieve relevant information, IR system has to understand user requests through query. The languages used to create a query are usually constrained and might not satisfy the normal rules of syntax. The query is refined through search statement where a representation of user query is formulated. The similarity computation uses this representative to match the given query by a user and those information that user would like to retrieve in response to the query. Various techniques can be used in similarity computation, usually by applying approximate match, which presenting different ranked lists of relevant document.

The process to match user query and information sources not only requires query analysis but also the analysis and representation of information. The key concept in IR is that the information to be retrieved is organised in a recoverable form by creating information representation. The analysis of the original information, might be as simple as identifying the title or as complex as performing linguistics analysis, is requires to create information representation. Indexing is one of the processes that can be used to analyse the information by creating an index. Basically, it contained selected term with the locations which the term occurred. In addition, this index adds an extra attributes for each term by assigning a weight value which is useful for search and retrieval process.

1.1.1 Contextual Retrieval

Workshop report on *Challenges in Information Retrieval and Language Modelling* outlined two long term challenges in IR research (James, Aslam et al. 2003). One of the challenges is contextual retrieval. Contextual retrieval is defined to combine three elements into single framework towards providing the most accurate answer for user's information needs. These three elements are search technologies, the understanding about query and the integration of user context. The contextual retrieval emerges because current IR system could not give the appropriate document as the user's want. Currently, the IR system treats every query submitted in isolation where each terms in query are judged on its own, and the context behind each request is usually ignored by IR system. In addition, the ranked list return by the IR system only takes the possibility of how this particular document is matches with the user's query. The representation of user preferences, search context or the task context is not taken into account.

1.1.2 Data Fusion

Data fusion is an approach which combines different data that comes from various sources, yet still refers to the same objects, in attempt to enhance the quality of output regarding the objects (Ng and Kantor 1998). This approach is implemented in IR areas since the principle "two heads are better than one" is believed could produce a better performance. Ever since it's implemented, the outcome by combining retrieval results shown consistence improvement (Smeaton 1998). In IR, data fusion

works by fusing two or more outputs i.e. ranked list of relevant documents that produces by dissimilar IR schemes or by using different query that expressing same information need and executed in single retrieval strategies. Much works has been done in this area especially in merging calculation and performance prediction (Nuraya and Can 2006; Xu and Benaroch 2006).

1.2 Problem Statement

The search and retrieval process in IR requires user queries. Query is an essential part in the whole IR system where its enable users to express their information need through it. Unfortunately, users often fail to state clearly their actual information need in the query form. This circumstance occurs when query often formulated with a simple and short word which presents some ambiguity (White, Jose et al. 2002; Shen, Tan et al. 2005(a)(b)). In addition, the IR system process works based on "Quality In, Quality Out" principle. Therefore, the more precise the query, the more relevant documents to the user's information need would be retrieved.

Beyond the ambiguity of the query, frequently the set of relevant documents retrieved by dissimilar IR systems is different to each other (Lee 1997; McCabe, Chowdhury et al. 2001; Tsikrika and Lalmas 2001). The availability of various techniques to compute the similarity computation between query and document contributes to this problem. Although these schemes retrieved different sets of relevant document, the accuracy in retrieving precisely relevant document is still questionable. Furthermore, the user does not have a time to submit a query to every