An Architecture to Enhance Post Retrieval Document Relevancy Using Integrated Techniques

Vimala Balakrishnan and Kian Ahmadi

***Abstract*—This paper is a proposal based on an on-going work in which we attempt to take advantage of information retrieval (IR) and case-based reasoning (CBR) techniques combined with the aim to improve the document relevancy of a search result. The proposed architecture contains two main phases: first is the IR phase whereby relevance feedback (RF) is implemented on search results produced based on adjacency keyword algorithm. Second is the CBR phase which further improves the results based on the output from phase one. This paper presents an explanation on the proposed RF-CBR model. It is believed that the integration of these two popular techniques would result in an improved document relevancy.**

***Index Terms*—Adjacency keywords, case-based reasoning, KNN algorithm, relevance feedback.**

I. Introduction

With ever growing information over the Web, finding high quality relevant information within the large collection of texts is a challenging issue. Traditional searching for information relies on exact match or "one-fits-all" principle search mechanism based on the use of Boolean queries and keywords. Such an approach often results in the same documents being returned to the users whenever the same keywords (regardless of the order) are used in the queries. Additionally this also means an overwhelming number of results are returned to the users to select [1]. This prompted researchers to explore best match mechanism, which relies on unstructured queries and ranks the search documents based on their relevancy [2].

The case-based reasoning (CBR) is based on the best match mechanism. It works on the basis of suggesting or solving new problems by adapting previous solutions to the new problems. The CBR paradigm covers a range of different methods for organizing, retrieving, utilizing and indexing the knowledge retained in past cases and it favors learning from experience, since it is usually easier to learn by retaining a concrete problem solving experience [3]. CBR has been widely and successfully used in medical domains [4], however recent studies have attempted to utilize this technique in the field of information retrieval (IR) as well [5].

By considering the importance of document relevance score in the ranked retrieval results and the advantages of CBR technique, this paper aims to propose an architecture that combines IR and CBR to improve document relevancy. The two main objectives are to improve IR mechanism by: *first,* maximizing the relevancy of the information retrieval system outputs to those intended by the user(s), and *second*, enhancing the similarities of the information retrieval outputs of the system's ranking to the human ranking using CBR.

The rest of the paper is organized as follows: the following section provides a brief description on IR, IR models, followed by previous works on relevance feedback (RF) and CBR. Then, the proposed architecture is presented and discussed. Future work and conclusion concludes the paper.

Manuscript received February 12, 2013; revise April 18, 2013.

The authors are with the Faculty of Computer Science and Information Systems, University of Malaya, Kuala Lumpur, Malaysia (e-mail: vimala.balakrishnan@um.edu.my, kian_santa@yahoo.com).

II. Information Retrieval

Information retrieval (IR) is the study of techniques of providing solution for retrieving text documents from storage relevant to the user needs. According to Manning et al. [6], IR is defined as "*ways to find materials (usually documents) that contain large number of texts (of an unstructured nature) that meet the information need from within large collections (usually stored on computers).*" IR approach is to find the items that are relevant and considered the best matches among the partial matching to the keywords defined in user query. It provides a solution with the best match results that rank to the degree of relevance in response to the user query [7]. A typical example of IR is the web search engine such as Google Search that is designed to search for information on the Web.

III. Information Retrieval Models

There are three classic IR models: *First*, the Boolean retrieval model which is based on set theory and Boolean algebra that model documents and queries as sets of index terms [8]. This is one of the earliest and simplest model in which keywords are logically combined with Boolean operators AND, OR and NOT to form the query in the retrieval system [8]. Although this retrieval approach is used in many commercial systems, but the drawbacks are well-known. For example, based upon two Boolean values of "zero" and "one", the retrieval results tend to have the effect of large number of documents or none at all, and thus it is difficult to control the output size [1]. Furthermore, Boolean retrieval return matching documents without taking relevance of the documents into consideration. This results in the user having to browse through the list to find the one that meets his or her information's need.

Second is the probabilistic retrieval model based on probability theory that ranks documents according to the probability of relevance [9]. In other words, for a given user

DOI: 10.7763/IJMLC.2013.V3.309

query, the documents in the collection that would most probably be relevant to the user is retrieved and then ranked according to the probability of relevance. The solution is based on the assumption that for each query, a document collection can be divided into two exclusive sets: the set of relevant and irrelevant documents. Therefore, every document in the collection can only be found in either one of these two sets. Some recent examples of studies that have explored or improved this theory include [10] and [11].

Third, is the vector space model (VSM) which models documents and queries as vectors in a multi-dimensional space [8]. A document vector is modeled as a list of index terms extracted from the document with associated weights representing the importance of the terms in the document [12]. Similarly, a query vector consists of a list of keywords with associated weights representing the importance of the keywords in the query [12]. The model is not without its limitations, for instance it assumes all terms are independently represented and related to each other only if the words are matched in the query and the document [13]. Many enhancements to VSM have been developed in the past including the study by Tai *et al.*[13] based on RF, multi-term VSM [14] based on adjacency terms relationship, and multi-term VSM based on adjacency keyword-order [15]. The current study extends the work by Lim *et al.* [15] by including RF and CBR.

The objective of any IR model is to find as many relevant documents as possible and at the same time retrieve a few possible non-relevant documents. Two common metrics used to evaluate the effectiveness of IR are precision and recall. Recall measures the number of relevant results that the system was able to retrieve, whilst precision measures the number of retrieved results that are actually relevant [12]. In IR, effectiveness is achieved when both these metrics are maximized, but this is difficult as in most instances, precision decreases as recall increases. When recall cannot be sacrificed, then increased average precision which is a score that considers the order in which relevant and non-relevant hits are presented to the users is used [5]. The goal of the paper is to improve document relevancy, using both IR and CBR techniques. Therefore, in this scenario, it is believed recall and precision are both required to determine the relevancy, therefore, average precision will be used. Various techniques exist to improve the average precision, such as relevance ranking algorithms, meta-searching, relevance feedback (RF) systems, etc. The undertaken study focuses on RF.

IV. Relevance Feedback

The effectiveness of a retrieval system may be expressed in terms of the relevance weighting focusing on improving the document scores, whereby the higher the score value, the more relevant a document is [8]. Two main approaches to RF include implicit and explicit RF. Explicit RF requires the users to explicitly rank documents to state their relevance, whereas implicit RF attempts to estimate relevancy based on user's behaviour such as amount of time spent on a site or mouse-click pattern. Studies have attempted to exploit these approaches in various ways. For instance, combining both implicit and explicit RF to improve image retrievals [16] and to recommend e-books [17]. Studies particularly focusing on explicit RF were carried out by Schefels et al. [18] to manage website visitor feedback and Takimoto [19] in the field of linguistic. Similarly, the current study adopts explicit RF via user ratings.

V. Case-Based Reasoning (CBR)

Case-based reasoning (CBR) is a four-step process used widely in computer reasoning [3]. First is the retrieve process which retrieves cases from the case-base to solve a problem. Second is the reuse step which maps the solution from the previous case to the target problem. Third is revise whereby the new solution is tested in the real world or a simulation , and finally fourth is retain whereby the adapted solution is stored as a new case-base [20]. Studies specifically emphasizing the use of CBR to improve IR include [21] in knowledge management system and [5] in digital forensics.

VI. Proposed Architecture

The RF-CBR integrated architecture can be illustrated as shown in Fig. 1. The two main phases, that is, RF and CBR are highlighted as well.

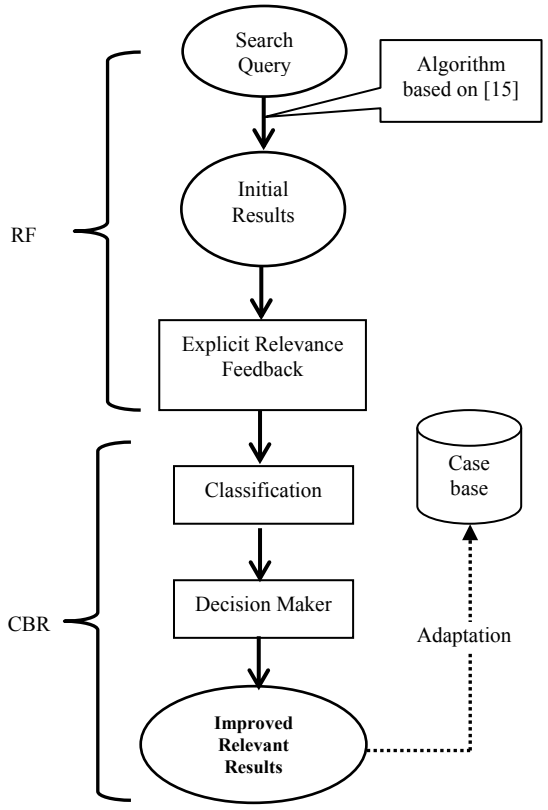

Fig. 1. RF-CBR integrated architecture

A. Phase One – Relevance Feedback

In this phase, two main techniques are employed. When a user performs a search query, the system will present an initial set of results using the algorithm introduced by Lim et al. [15]. The previous work is based on the concept of keyword grouping, whereby the keyword-order relationship

in the adjacency terms is used to measure a term's weight, and thus document relevancy are improved [15]. Next would be the explicit RF technique, in which the system will acquire feedback from the users who will determine the relevancy of the documents by providing ratings. These ratings will then be averaged and sorted so that highly relevant documents are identified. These outputs will be fed into the next phase.

B. Phase Two - Case-Based Reasoning

Classification - Using the output from RF, this module searches the past cases and picks the most relevant matches from the case base using the weighted K-nearest neighbour (KNN) algorithm, which stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). The study adopts the Hamming distance function to calculate the similarity between cases considering the fact that explicit RF (i.e. ratings) would be discrete in nature (refer to [22] for further details on KNN algorithm).

Decision maker - It takes as input the cases retrieved by the classifier. These are then analyzed to retrieve the most relevant documents to be returned to the users. The relevancy is then adapted to the case-base if it is new. The case-base archives the relevancy rules or patterns that were previously used, and these can be adapted to solve other similar cases in future.

VII. Future Work

The next stage of the current study would be the implementation process, followed by the evaluation. The prototype system would be tested against the academic journal collections provided by TREC. Additionally, the performance of RF-CBR will also be compared against other techniques, for instance, using only explicit RF or only CBR. This would then provide us with the knowledge if the hybrid mechanism works better than stand-alone techniques.

VIII. Conclusion

With the growth of Internet, information became widely available to people, hence studies on information retrieval (IR) became popular. The main focus of most of the studies was to improve document relevancy so that users do not waste unnecessary time searching for documents that are relevant to them. Relevance feedback is one of the common techniques used to improve relevancy based on feedback (both explicit and implicit), whereas case-based reasoning (CBR) works on the basis of best match algorithm. The current study aims to improve document relevancy by integrating RF and CBR. The architecture is proposed and briefly described in this paper. It is believed that with the successful implementation of this proposal, document relevancy can be further improved.

References

[1] S. Agrawal, S. Chaudhuri, and G. Das, "DBXplorer: a system for keyword-based search over relational databases," in *Proc. 18th International Conference on Data Engineering,* pp. 5 –16.2002.

[2] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: two sides of the same coin?" *Communication in ACM*, vol. 35, no. 12, pp. 29–38, Dec. 1992.

[3] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *Artificial Intelligence Communications,* vol. 7, no. 1, pp. 39–59, Jan. 1994.

[4] V. Balakrishnan, M. R. Shakouri, and H. Hoodeh, "Integrating Association Rules and Case-based Reasoning to Predict Retinopathy," *Maejo International Journal of Science and Technology*, vol. 6, no. 3, pp. 334-343, 2012.

[5] N. L. Beebe, J. G. Clark, G. B. Dietrich, M. S. Ko, and D. Ko, "Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies," *Decision Support Systems*, vol. 51, no. 4, pp. 732–744, Nov. 2011.

[6] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval,* vol. 1. Cambridge University Press Cambridge, 2008.

[7] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval,* vol. 463. ACM press New York, 1999.

[8] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Communication of ACM*, vol. 26, no. 11, pp. 1022–1036, Nov. 1983.

[9] S. Chaudhuri, G. Das, V. Hristidis, and G. Weikum, "Probabilistic ranking of database query results," in *Proceedings of the international conference on very large data bases*, 2004, pp. 888–899.

[10] S. Bansal and R. Garg, "A Novel Probabilistic Approach for Efficient Information Retrieval," *International Journal of Computer Applications,* vol. 9, no. 2, pp. 44–48, Nov. 2010.

[11] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski, "Probabilistic retrieval and visualization of biologically relevant microarray experiments," *Bioinformatics*, vol. 25, no. 12, pp. i145–i153, Jun. 2009.

[12] D. L. Lee, H. Chuang, and K. Seamons, "Document Ranking and the Vector-Space Model," *IEEE Softw.*, vol. 14, no. 2, pp. 67–75, Mar. 1997.

[13] X. Tai, F. Ren, and K. Kita, "An information retrieval model based on vector space method by supervised learning," *Information Processing & Management,* vol. 38, no. 6, pp. 749–764, Nov. 2002.

[14] L. S. Wang, "Relevance weighting of multi-term queries for Vector Space Model," in *Proc. IEEE Symposium on Computational Intelligence and Data Mining, 2009. CIDM '09*, pp. 396 –402, 2009.

[15] B. H. Lim, V. Balakrishnan, and R. G. Raj, "Improving the Relevancy of Document Search using the Multi-Term Adjacency Keyword-Order Model," *Malaysian Journal of Computer Science*, vol. 25, no. 1, pp. 1–10, 2012.

[16] G. Raghuwanshi, N. Mishra, and S. Sharma, "Content based Image Retrieval using Implicit and Explicit Feedback with Interactive Genetic Algorithm," *International Journal of Computer Applications*, vol. 43, no. 16, pp. 8–14, Apr. 2012.

[17] E. R. Núñez-Valdéz, J. M. Cueva Lovelle, O. Sanjuán Martínez, V. García-Díaz, P. Ordoñez de Pablos, and C. E. Montenegro Marín, "Implicit feedback techniques on recommender systems applied to electronic books," *Computers in Human Behavior*, vol. 28, no. 4, pp. 1186–1193, Jul. 2012.

[18] C. Schefels and R. V. Zicari, "A framework analysis for managing feedback of visitors of a web site," *International Journal of Web Information Systems,* vol. 8, no. 1, pp. 127–150, Mar. 2012.

[19] M. Takimoto, "The effects of explicit feedback on the development of pragmatic proficiency," *Language Teaching Research*, vol. 10, no. 4, pp. 393–417, Oct. 2006.

[20] K. A. Kumar, Y. Singh, and S. Sanyal, "Hybrid approach using case-based reasoning and rule-based reasoning for domain independent clinical decision support in ICU," *Expert Systems with Applications*, vol. 36, no. 1, pp. 65–71, Jan. 2009.

[21] N. Gronau and F. Laskowski, "Using Case-Based Reasoning to Improve Information Retrieval in Knowledge Management Systems," in *Advances in Web Intelligence*, vol. 2663, E. Menasalvas, J. Segovia, and P. S. Szczepaniak, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 94–102.

[22] R.O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd Edition, Wiley International - Google Scholar, 2000.

V. Balakrishnan was received her PhD in the field of Ergonomics in 2009 from Multimedia University, Malaysia. Both her Masters and Bachelor degrees were from University of Science, Malaysia.

She is currently affiliated with the Faculty of Computer Science and Information Technology, University of Malaya as a Senior Lecturer. Most of her research works are in the field of data engineering, opinion mining, information retrieval and health informatics.

Dr. Balakrishnan is also a member of the Medical Research Support (Medicres) group and Global Science and Technology Forum.

K. Ahmadi earned his Degree in Computer Science IT Management from University of Malaya, Malaysia in 2011. He is currently doing his Master in Computer Science in University of Malaya, Malaysia.

He worked as a technical administrator and technical assistant (2000-2006) and research assistant in University of Malaya (2010-2011). His research interest fields are Wireless networks and Networking System, Intelligent System and Information System.

Mr. Ahmadi was a member of IEEE Journal of Computer Science (2010-2011). He has two patents under his name: *"The Intelligent Vehicular Communication Protocol (Inv.Comm) Based on driving safety & traffic efficiency Improves emergency cases", PI 2010001264,* Malaysia 2010 *and "Novel Assistive Software Application (NASA) Helping deaf-mute people by using mobile phone and PDA",* Malaysia (under process). Also he won a bronze medal in Malaysia Technology Expo MTE 2011.