The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1741-0398.htm

Received 22 August 2010
Revised 7 December 2010
13 February 2011
Accepted 28 February 2011

A combined measure for representative information retrieval in enterprise information systems

Baojun Ma, Qiang Wei and Guoqing Chen
School of Economics and Management, Tsinghua University, Beijing, China

Abstract

Purpose – The purpose of this paper is to propose a framework for describing and evaluating the representativeness of a small set of search results extracted from the original results: this is deemed desirable in information retrieval in enterprise information systems.

Design/methodology/approach – The paper proposes a combined measure, namely RF_β, to evaluate the extracted small set in terms of the notions of coverage and redundancy. Data experiments were conducted on three different extraction strategies to evaluate the representativeness, i.e. coverage and redundancy.

Findings – Both from intuitive and experimental perspectives, the proposed coverage measure, redundancy measure and RF_β measure could effectively evaluate the representativeness.

Research limitations/implications – The search results, e.g. in the form of documents and texts, are modeled using a vector space model and cosine similarity. Semantic models and linguistic models could be further introduced into this research to improve the proposed measures.

Practical implications – With the rapidly growing need for information retrieval in enterprise information systems, the representativeness of search results become more desirable and important for search engine users. The well-designed representativeness measures will help them achieve satisfactory results.

Originality/value – The originality of the paper lies in the definition of representativeness of a small set of search results extracted from the original results. This focuses on the two aspects of coverage rate and redundancy rate both from intuitive and experimental perspectives.

Keywords Information retrieval, Representativeness, Coverage, Redundancy, Set theory, Information systems

Paper type Research paper

1. Introduction

Information retrieval (IR) (Liu, 2007), as one of the key technologies adopted by search engines (Spink and Jansen, 2004), not only is widely used for public web search, but also plays a more important role in enterprise information systems (Hawking, 2004; Balog, 2007; Broder and Ciccolo, 2004). In enterprise information management, the major challenge faced by information retrieval is to effectively and efficiently organize huge amounts of data from multiple sources, e.g. internet, intranet, file systems,

Journal of Enterprise Information Management
Vol. 24 No. 4, 2011
pp. 310-321

1741-0398
DOI 10.1108/17410391111148567

The work was partly supported by the National Natural Science Foundation of China (70890080/71072015), the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities of China (7JJD63005), and Tsinghua-Leuven Research Cooperation Project (3H051154).

document management systems, OA system, internal blogs, transaction databases, etc., and then present a consolidated list of quality-ranked contents, e.g. records and web pages, from these various sources (Hawking, 2004; Broder and Ciccolo, 2004).

In real-world information retrieval applications in enterprises, a search engine may often generate a huge volume, e.g. in size of thousands of search results (e.g. web pages, documents, texts, records), all of which satisfy the search criteria but could hardly be browsed one-by-one by users. Usually, users would prefer a small set of search results that appear in the first couple of web pages and have a good quality of the search. This is deemed meaningful and important to search engine users, as most users do not seem to be interested (or become much less interested) in browsing the search results that are displayed in later pages. Particularly in a mobile search environment (e.g. with cell phones) that is getting popular and pervasive nowadays in businesses and social lives, browsing many later pages is neither interesting nor practical. Thus, the quality of a small set of results to be top-ranked and displayed to users is of great interest to both academia and practitioners.

Nevertheless, since different users may have discrepant search requirements even with same keywords, different criteria of are used for evaluating the quality of IR. Many quality measures are concerned with the relevance between keywords and search results (Zhu *et al.*, 2009), such as recall, precision (Kraft and Bookstein, 1978) and some combined measures (e.g. *F*-value (van Rijsbergen, 1979), *R*-precision, MAP (Buckley and Voorhees, 2000), NDCG (Sakai, 2007) and so on (Manning *et al.*, 2009)). However, given a specific keyword, which is quite short normally, all of the search results are highly relevant to the keyword, i.e. containing the keyword, which means that relevance (e.g. precision) is not a problem, but cannot be used to effectively filter out a small set for users. Thereafter, some other quality measures should be considered, e.g. PageRank measure (Page *et al.*, 1998), etc., to further rank the search results by their "importance" or "hotness". Moreover, with top-*k* (ranking) extraction method, which is widely used in web search and information retrieval (Guntzer *et al.*, 2000; Bruno *et al.*, 2002; Fagin *et al.*, 2003; Ilyas *et al.*, 2008; Mamoulis *et al.*, 2007; Marian *et al.*, 2004), the *k* search results with the highest values for certain ranking function from the original set could be obtained and presented to users. In general, the ranking function is based on some attributes of search results, e.g. the freshness of time, the number of visits, the number of comments, the "hotness", the "importance", etc. (Lian and Chen, 2009; Papadias *et al.*, 2005; Yiu and Mamoulis, 2009). From semantic and intuitive perspective, this method tries to obtain the most "important" search results.

Top-*k* extraction with some measures can present users with satisfactory, e.g. top-hotness-ranked, results in many cases. nevertheless, information retrieval in enterprise still face some new challenges. First, the top-ranked results, e.g. on the first several pages, cannot effectively reflect the overall information of all retrieved search results. In web search, it is generally not a severe problem, since users of web search usually prefer to get the basic information in the first piece of all search results. For search in enterprise information systems, however, users may sometimes prefer to get a clear view of the whole results satisfying the search keyword. For example, in a Wiki system of an enterprise, if a user wants to query and review some knowledge, he/she may prefer to get more overviewed information of all results relevant to the keyword, while the top-ranked results can only represent a very limited part of the overall information. Thus, how to evaluate the information covered by the small set referred the

whole set is worthy to analyze and can help extract a better representative set. Second, it is not seldom seen that, the top-ranked results are sometimes redundant in contents. This is because that multiple data sources in enterprise information systems may contain many similar and duplicated contents (Hawking, 2004; Balog, 2007; Broder and Ciccolo, 2004), and differentiation of results on content is a new challenge for search engine. In this case, users may be presented with results containing redundant contents, which not only decreases the quality of information retrieval but significantly impact users' search experience. Thereafter, how to evaluate the redundancy of a set of results is also one of the important aspects of information retrieval.

In this paper, we focus on the quality of search in enterprise information management, in light of representativeness, on two respects: one is the coverage measure referring to the number of the results in the original set "covered" by the extracted small set of results; the other is the redundancy measure referring to the number of "redundant" results in the extracted small set. These two measures reflect the representativeness of a small set referred to the original set in two different ways. Merely for illustrative purposes, let us first consider a simplified crisp case where the original collection of retrieved results is {A, B, C, C, D} with one extra duplicated C, and a small set with three results could be browsed by users. Given two derived small sets {A, B, C} and {A, C, C}, the former can cover four results and the latter can cover three results, respectively, whereas the latter is more redundant than the former. In this case, the collection {A, B, C} will be more suitable to be presented to users than {A, C, C}, since it can convey more information to users and is less redundant. Furthermore, for this example, {A, B, C, C, D} could be clustered into four sets, i.e. {A}, {B}, {C, C} and {D}, in which one result could be extracted to finally construct a new set, i.e. {A, B, C, D}. Thereafter, {A, B, C, D} can cover the overall information of {A, B, C, C, D} and is non-redundant, which is called a representative set.

In usual search cases, however, the search results (documents, web pages, texts, etc.) are generally close/similar to each other (e.g. via text similarity) on contents. It is easy to encounter a situation where some search results are almost identical or highly similar. This may be, on one hand, a reflection on a frequent attention to something, or on the other hand a reflection on a highly duplicate piece of information. Such an example can be a number of work documents, each being a cited report on the same act from a single source, which reads similar but is hardly interesting. In doing so, not only the coverage measure and redundancy measure should be further extended in the framework of text similarity, but fuzzy clustering methods with text similarities should also be conducted to help derive clusters, in which the results are quite similar to each other, to help extract the representative set with high coverage and low redundancy. As a matter of fact, a combined view of representativeness in coverage and redundancy is deemed desirable, which motivates the effort of this study.

This paper is organized as follows. Section 2 briefly introduces some related works on existed measures to evaluate the search quality and existed methods to extract a small set from the original set of search results. In section 3, the coverage measure and redundancy measure are defined and discussed with some important properties, based on which the combined RF_β measure is introduced. To evaluate the effectiveness of the proposed measures, some empirical experiments are conducted and discussed in section 4. Finally, some concluding remarks are presented in section 5.

2. Related works

Generally, high representativeness of a small set referred to the original set has the intuitive meaning that the small set covers high fraction of information of the large set while itself possesses little information redundancy.

To evaluate the coverage of a small set referred to an original set, researchers have done some works, e.g. subtopic recall at rank K (Zhai *et al.*, 2003), *Representative Coverage* (Pan *et al.*, 2005). To evaluate information redundancy between two search results, several measurements like Kullback-Leibler divergence (Pan *et al.*, 2005; Zhai *et al.*, 2003; Zhang *et al.*, 2002), cosine-similarity-type measure (Zhang *et al.*, 2002) and keyword number of the intersection of two keyword sets of documents (Zhang *et al.*, 2002), maximum similarity (Carbonell and Goldstein, 1998) have been proposed. Nevertheless, these works did not take coverage and redundancy into a unified framework, and could not be easily implanted into search engine in enterprise information management.

Normally, clustering methods are deemed effective to extract a representative set referred an original set (Liu, 2007; Han and Kamber, 2006). Generally speaking, clustering is a unsupervised classification of a given set of search results into clusters such that the results within each cluster are similar to each other, and the results from different clusters are dissimilar to each other (Aliguliyev, 2009; Carpineto *et al.*, 2009; Liu, 2007; Han and Kamber, 2006; Hastie *et al.*, 2001; Grabmeier and Rudolph, 2002; Jain *et al.*, 1999). Thus, with the original set of search results, given a clustering method, k clusters/sets could be derived, in each of which a search result with the maximum similarities to other results in the cluster could be extracted as the representative result, since the result can cover maximum information of all the results in the cluster. Moreover, since the k results are from k clusters respectively, where the similarities among different clusters are low, the k results will be mutually less redundant on content. Thereafter, the finally derived set of k representative results from k clusters respectively is called the representative set, which not only can cover the information of the original set to a large extent, but is low redundant.

There exist many types of clustering methods, e.g. k-means clustering methods, graph-based clustering methods, agglomerative based clustering methods (Aliguliyev, 2009; Carpineto *et al.*, 2009; Liu, 2007; Han and Kamber, 2006; Hastie *et al.*, 2001; Grabmeier and Rudolph, 2002; Jain *et al.*, 1999). With these methods, some representative set could extracted, which tend to be with high coverage and low redundancy. In the following discussions, a well-known k-means clustering method, i.e. Direct, is considered for evaluating the proposed coverage measure, redundancy measure, as well as a combined measure.

3. Evaluation measures of representativeness

Let us concentrate our discussion on the quality of a small set out of the set of overall search results. Given a set of n search results, e.g. $D = \{d_1, d_2, \ldots, d_n\}$, where d_i is a search result, $i = 1, \ldots, n$, then an IR method E is to extract k documents from D ($k \leq n$) under certain criteria, resulting in D^E, where $D^E \subseteq D$. This paper is to investigate the representativeness of D^E with respect to D on two aspects: the degree that D^E can cover the information of D (i.e. the coverage rate); and the degree of redundancy existing in D^E (i.e. the redundancy rate). Intuitively, given D and k, a good IR method is

to extract a D^E with as large coverage rate as possible and as small redundancy rate as possible.

3.1 Coverage measure

For two search results d and d', we call d is close to d' with degree $F_C(d, d') \in [0, 1]$, where $F_C(d, d')$ is the degree of similarity/closeness, which is reflexive and symmetric and can be calculated with cosine-similarity measure (Baeza-Yates and Ribeiro-Neto, 1999; Liu, 2007; Manning *et al.*, 2009; Salton, 1971). Thus, given two sets of documents D and D' and a document $d \in D$, D' is called to cover d with degree $= \max_{d' \in D'}(F_C(d', d))$. Moreover, the rate that D' covers D (i.e. the coverage rate $r_C(D', D)$) could be defined as follows:

$$r_C(D', D) = \sum_{d \in D} \left(\max_{d' \in D'} (F_C(d', d)) \right) / |D| \tag{1}$$

where $|D|$ is the number of documents in D.

For example, first consider a crisp case (i.e. $F_C(d', d) = 1$ if $d' = d$, otherwise 0). Suppose $D = \{A, B, C, C, D\}$, $D_1^E = \{A, B, C, C\}$ and $D_2^E = \{A, B, C\}$. Then we have $r_C(D_1^E, D) = 4/5$, and $r_C(D_2^E, D) = 4/5$, because both cover four documents in D. Furthermore, in a closeness case with the closeness measure $F_C(d', d) = |d' \cap d|/4$ (i.e. $|d' \cap d|$ represents the number of elements that d' and d share), suppose $D = \{ABCD, ABCE, FGHI, FGHI, FGHJ\}$, $D_1^E = \{ABCD, ABCE, FGHJ\}$ and $D_2^E = \{ABCE, FGHI\}$, we have $r_C(D_1^E, D) = 9/10$, and $r_C(D_2^E, D) = 9/10$ as well. Note that both D_1^E and D_2^E cover the same number of search results on contents. However, they have different levels of redundancy, which will be discussed in the next section.

Moreover, the coverage rate has the following properties:

- $0 \leq r_C(D', D) \leq 1$.
- If $D = D'$, $r_C(D', D) = 1$; if $D' = \varnothing$ and $D \neq \varnothing$, $r_C(D', D) = 0$.
- $r_C(D', D) \neq r_C(D, D')$, except for $D = D'$ or $|D| = |D'| = 1$.
- If $D' \subseteq D$, then $0 < r_C(D', D) \leq 1$ and $r_C(D, D') = 1$.
- If $D' \subseteq D$, then $r_C(D', D') \leq r_C(D, D')$.
- Denote $D - D' = \{d \mid d \in D \text{ and } d \notin D'\}$, if $D - D' \neq \varnothing$, then $r_C(D', D - D') \leq r_C(D', D)$.

where D, D' and D' are nonempty sets of search results.

3.2 Redundancy measure

For a set D and a document d, $d \in D$, the degree that d is redundant in D is $1 - 1/\Sigma_{d' \in D} F_C(d', d)$. Furthermore, the redundancy rate of D could be defined:

$$r_R(D) = \sum_{d \in D} \left(1 - 1 / \sum_{d' \in D} F_C(d', d) \right) / |D| \tag{2}$$

Referring to the above examples, in the crisp case, $r_R(D_1^E) = 1/4$ and $r_R(D_2^E) = 0$. In the closeness case, $r_R(D_1^E) = 2/7$ and $r_R(D_2^E) = 0$. In a combined view, in either case,

D_2^E is considered better in representativeness than D_1^E since D_2^E has the same degree of coverage as D_1^E but a lower degree of redundancy than D_1^E. Moreover, for any nonempty set D, the redundancy rate has the following properties:

- $0 \leq r_R(D) < 1$.
- $r_R(D) = 0$, if $|D| = 1$.
- $r_R(D) = 0$, if $F_C(\mathrm{d}, d') = 0$, $\forall d, d' \in D, d \neq d', |D| > 1$.
- $r_R(D) = 1 - 1/|D|$, if $F_C(d, d') = 1$, $\forall d, d' \in D, d \neq d'$.

If a search result d, $d \in$ D, is with $1 - 1/\Sigma_{d' \in D} F_C(d', d) > r_R(D)$, then $r_R(D - \{d\}) < r_R(D)$.

3.3 RF_β measure combining coverage and redundancy

As discussed in the previous section, high representativeness means high coverage and low redundancy. Hence, a combined view is regarded necessary. In sprit of recall, precision and F_β (Kraft and Bookstein, 1978; van Rijsbergen, 1979), a combined measure, namely RF_β, could be defined as follows:

$$RF_\beta(D', D) = \frac{1}{\alpha / r_C(D', D) + (1 - \alpha) / (1 - r_R(D'))} = \frac{(\beta^2 + 1) r_C(D', D) \times (1 - r_R(D'))}{\beta^2 \times r_C(D', D) + (1 - r_R(D'))} \quad (3)$$

Where $\beta^2 = (1 - \alpha)/\alpha$, $\alpha \in [0, 1]$, $\beta \in [0, +\infty)$. $RF_\beta(D', D)$ is a weighted harmonic mean of coverage rate and redundancy rate, where α or β reflects users' preference on coverage and non-redundancy. If $0 \leq \alpha < 0.5$ ($\beta > 1$), it means that users prefer more on non-redundancy than coverage, and if $0.5 < \alpha \leq 1$ ($0 \leq \beta < 1$), it means that users prefer more on coverage than non-redundancy. If $\alpha = 0.5$ ($\beta = 1$), it means that user treats coverage and non-redundancy equally. In addition, we have:

- $0 \leq RF_\beta(D', D) \leq 1$.
- Given a certain α (β), $RF_\beta(D', D)$ increases monotonously with $r_C(D', D)$'s increase and decreases monotonously with $r_R(D')$'s increase.

Take the same example as shown in the previous section, with $\alpha = 0.5$ ($\beta = 1$), in the crisp case, we have $RF_\beta(D_1^E, D) = 24/31 < 8/9 = RF_\beta(D_2^E, D)$, while in the closeness case, $RF_\beta(D_1^E, D) = 90/113 < 18/19 = RF_\beta(D_2^E, D)$, which conforms to the fact that D_2^E has a higher level of representativeness that D_1^E.

4. Empirical data experiments

In order to verify the coverage measure and redundancy measure, some empirical data experiments were conducted to compare the top-k search results of Google search engine, i.e. Google-k strategy, with the k representative results by clustering method, i.e. Clustering strategy, as well as k results selected using Random extraction strategy, i.e. Random strategy.

Usually, Google provides (in display) around 1,000 result items relevant to query keywords (though the total number of the results (e.g. millions of items) is often

indicated), which can be regarded as the original set D. However, users normally only browse the first several pages, e.g. k search results ($k \ll 1{,}000$), to search their preferred documents. Though Google's first k search results, i.e. with Google-k strategy, were with high PageRank values and might have been diversified considering similarity, many search results were still found quite similar, e.g. the hottest content relevant to keywords usually appears frequently in different search results, which may imply high information coverage but high redundancy. In Clustering strategy, the Direct method is used to cluster D into k information-equivalent classes with k-means methodology and extract one representative search result for each class, which tries to obtain low redundancy without significant loss of information coverage (Liu, 2007). Moreover, the CLUTO software package is selected to perform the Direct clustering and extraction (Zhao and Karypis, 2004; Zhao and Karypis, 2005). Furthermore, a random extraction strategy (hereafter called Random) is to randomly extract k documents in D with uniform distribution. Moreover, the cosine similarity measure in the vector space IR model (Salton, 1971; Manning *et al.*, 2009) is used to obtain the degree of similarity between web documents.

In the empirical experiments, for comparison purpose, the benchmark data in the KDD Cup 2005 task is selected (Li *et al.*, 2005), which is widely used in performance evaluation in information retrieval. Thereafter, the 111 queries provided by KDD Cup 2005 data are chosen as the search keywords in Google. For comparison, the experiments were conducted with extraction size $k = 10$, 20, and 30, respectively, which approximately represent one, two, and three web pages of search results. For Random strategy, in order to narrow the deviations, the listed values are the means of 50 IID (Independently Identically Distributed) extractions. Note that we conducted the experiments on a 3.00 GHz 2.96 Gb RAM machine running Microsoft Windows XP Professional, and used Java language. For obtaining and analyzing the contents of web pages provided by Google, we used Apache Lucene, HTML parser and http client packages and APIs.

Figures 1 to 3 show the coverage rates and redundancy rates of extracted search results by Google-k, Clustering and Random, with $k = 10$, 20 and 30, respectively.

Figures 1 to 3 show that, first, the coverage rates of Clustering were the highest and its redundancy rates were the lowest in most cases, meaning that Clustering could

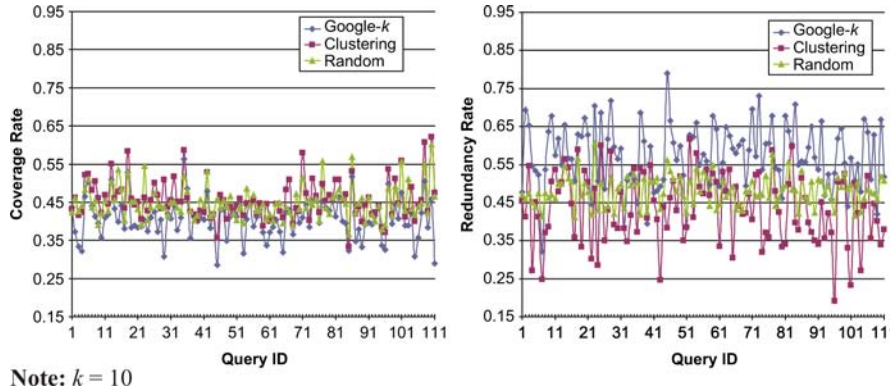
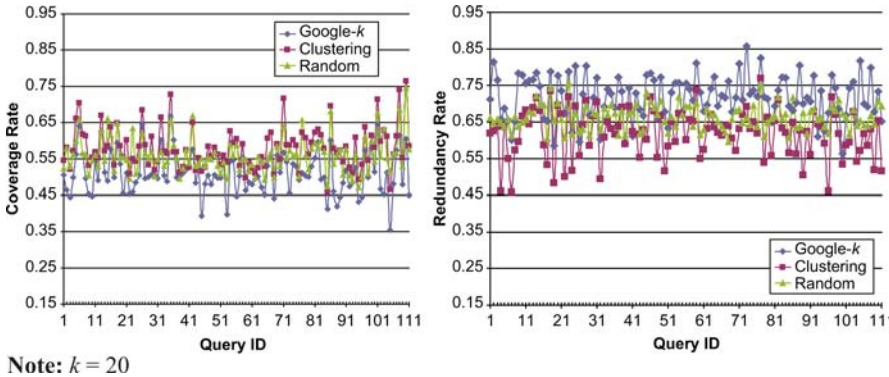
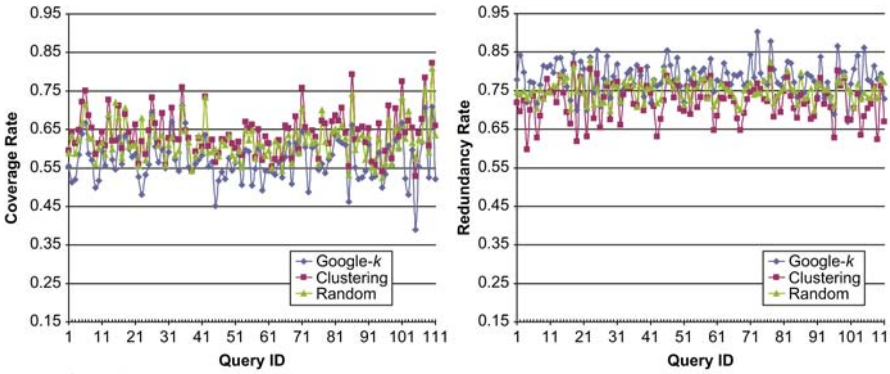


Figure 1.
The coverage rates and redundancy rates ($k = 10$)



Note: $k = 20$

Figure 2.
The coverage rates and redundancy rates ($k = 20$)



Note: $k = 30$

Figure 3.
The coverage rates and redundancy rates ($k = 30$)

extract representative search results effectively, specifically on information coverage and redundancy. Second, roughly, both on coverage rate and redundancy rate, Google-*k* seems to perform the worst compared with Clustering and Random strategies (i.e. the lowest coverage rates and highest redundancy rates in most cases), while Random presents a neutral performance both on information coverage and redundancy, which is also consistent with intuition. Further, with the increase of k, both of these two rates of the three strategies would keep increase, since the larger the subset is, the higher possibility to cover more information and generate more redundancy, which are also consistent to the properties discussed in Section 3.

To retrieve a more reliable analysis on the evaluation of coverage measure and redundancy measure on the three strategies, i.e. Google, Clustering and Random. A paired *t*-test is conducted on the empirical data experiments on 111 queries with $k = 10$, 20 and 30. Table I shows the statistical analysis results.

The statistical results in Table I show that, coverage rates of Clustering strategy is significantly higher than those of Google and Random strategies, while redundancy rates of Clustering strategy are significantly lower than those of Google and Random strategies, which are consistent with the theoretical analysis and above experimental results.

Extraction size	Measures	Assumptions	*t*-test value	Significance
$k = 10$	Coverage	Clustering > Google-k	13.758	***
		Clustering > Random	2.218	*
	Redundancy	Clustering < Google-k	16.196	***
		Clustering < Random	5.856	***
$k = 20$	Coverage	Clustering > Google-k	19.745	***
		Clustering > Random	7.227	***
	Redundancy	Clustering < Google-k	20.786	***
		Clustering < Random	8.502	***
$k = 30$	Coverage	Clustering > Google-k	19.942	***
		Clustering > Random	8.361	***
	Redundancy	Clustering < Google-k	18.851	***
		Clustering < Random	8.903	***

Notes: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table I.
Paired *t*-test on coverage rates and redundancy rates ($k = 10$, 20 and 30)

Additionally, by setting $\beta = 0$ (100% preference to coverage), 0.5, 1 (equal preference to coverage and non-redundancy), 2, 10 and 100 (almost 100 per cent preference to non-redundancy), more experiments were conducted to further examine the correspondingly RF_β values ($k = 10,\ 20,\ 30$), which were also consistent with Figures 1 and 3. Therefore, as discussed previously, the proposed coverage measure, redundancy measure, as well as RF_β measure, could help effectively evaluate the quality of IR in light of combining users' preferences on information coverage and redundancy in search results.

5. Conclusion and future work

This paper has proposed a representativeness measure RF_β to consider two concerns relating to the extracted small search set, i.e. coverage and redundancy, in a combined manner. Theoretical analysis shows that the proposed coverage measure and redundancy measure, as well as the combined RF_β measure can effectively evaluate the quality of extracted set referred to a given original set of search results. Empirical experiments with the benchmark data were conducted to compare three IR strategies, namely Google-k, Clustering and Random, verifying the effectiveness of the proposed measures, which are also consistent with theoretical analysis. Future studies could center on constructing a novel IR method for extracting representative information based on the RF_β measure and conducting case study on applications in enterprise information management.

References

Aliguliyev, R.M. (2009), "Clustering of document collection – a weighting approach", *Expert Systems with Applications*, Vol. 36 No. 4, pp. 7904-16.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM Press, New York, NY.

Balog, K. (2007), "People search in the enterprise", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands*, p. 916.

Broder, A.Z. and Ciccolo, A.C. (2004), "Towards the next generation of enterprise search technology", *IBM Systems Journal*, Vol. 43 No. 3, pp. 451-4.

Bruno, N., Chaudhri, S. and Gravand, L. (2002), "Top-*k* selection queries over relational databases", *ACM Transactions on Database Systems*, Vol. 27 No. 2, pp. 153-87.

Buckley, C. and Voorhees, E.M. (2000), "Evaluating evaluation measure stability", *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and development in Information Retrieval*, ACM Press, New York, NY, pp. 33-40.

Carbonell, J. and Goldstein, J. (1998), "The use of MMR, diversity-based reranking for reordering documents and producing summaries", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, NY, pp. 335-6.

Carpineto, C., Osinski, S., Romano, G. and Weiss, D. (2009), "A survey of web clustering engines", *ACM Computing Surveys*, Vol. 41 No. 3, p. 17.

Fagin, R., Lotem, A. and Naor, M. (2003), "Optimal aggregation algorithms for middleware", *Journal of Computer and System Sciences*, Vol. 66 No. 4, pp. 614-56.

Grabmeier, J. and Rudolph, A. (2002), "Techniques of cluster algorithms in data mining", *Data Mining and Knowledge Discovery*, Vol. 6 No. 4, pp. 303-60.

Guntzer, U., Balke, W. and Kießling, W. (2000), "Optimizing multi-feature queries for image databases", *Proceedings of the 26th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 419-28.

Han, J.W. and Kamber, M. (2006), *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufman Publishers, San Francisco, CA.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, Springer-Verlag, New York, NY.

Hawking, D. (2004), *Challenges in Enterprise Search, ACM International Conference Proceedings Series, Dunedin, New Zealand*, Vol. 52, pp. 15-24.

Ilyas, I.F., Beskales, G. and Soliman, M.A. (2008), "Survey of top-k query processing techniques in relational database systems", *ACM Computing Surveys*, Vol. 40 No. 4.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999), "Data clustering: a review", *ACM Computing Surveys*, Vol. 31 No. 3, pp. 264-323.

Kraft, D.E. and Bookstein, A. (1978), "Evaluation of information retrieval system: a decision theory approach", *Journal of the American Society for Information Science*, Vol. 29 No. 1, pp. 31-40.

Li, Y., Zheng, Z. and Dai, H. (2005), "KDD CUP-2005 report: facing a great challenge", *SIGKDD Explorations*, Vol. 7 No. 2, pp. 91-9.

Lian, X. and Chen, L. (2009), "Top-*k* dominating queries in uncertain databases", *Proceedings of the 12th International Conference on Extending Database Technology*, ACM Press, New York, NY, pp. 660-71.

Liu, B. (2007), *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, Berlin, Heidelberg, New York, NY.

Mamoulis, N., Cheng, K.H., Yiu, M.L. and Cheung, D.W. (2007), "Efficient top-*k* aggregation of ranked inputs", *ACM Transactions on Database Systems*, Vol. 32 No. 3, Article 19.

Manning, C.D., Raghavan, P. and Schütze, H. (2009), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge.

Marian, A., Bruno, N. and Gravano, L. (2004), "Evaluating top-*k* queries over web-accessible databases", *ACM Transactions on Database Systems*, Vol. 29 No. 2, pp. 319-62.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998), *The Pagerank Citation Ranking: Bringing Order to the Web, Technical Report*, Standford InfoLab.

Pan, F., Wang, W., Tung, A.K.H. and Yang, J. (2005), "Finding representative set from massive data", *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)*, IEEE Computer Society, NW Washington, DC, pp. 338-45.

Papadias, D., Tao, Y., Fu, G. and Seeger, B. (2005), "Progressive skyline computation in database systems", *ACM Transactions on Database Systems*, Vol. 30 No. 1, pp. 41-82.

Sakai, T. (2007), "On the reliability of information retrieval metrics based on graded relevance", *Information Processing & Management*, Vol. 43 No. 2, pp. 531-48.

Salton, G. (1971), *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice Hall, Englewood Cliffs, NJ.

Spink, A. and Jansen, B.J. (2004), *Web Search: Public Searching of the Web*, Kluwer Academic Publishers, New York, NY, Boston, MA, Dordrecht, London, Moscow.

van Rijsbergen, C.J. (1979), *Information Retrieval*, Butterworths, London.

Yiu, M.L. and Mamoulis, N. (2009), "Multi-dimensional top-*k* dominating queries", *The VLDB Journal*, Vol. 18 No. 3, pp. 695-718.

Zhai, C.X., Cohen, W.W. and Lafferty, J. (2003), "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, NY, pp. 10-17.

Zhang, Y., Callan, J. and Minka, T. (2002), "Novelty and redundancy detection in adaptive filtering", *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*, Vol. 02, ACM Press, New York, NY, pp. 81-8.

Zhao, Y. and Karypis, G. (2004), "Empirical and theoretical comparisons of selected criterion functions for document clustering", *Machine Learning*, Vol. 55 No. 3, pp. 311-31.

Zhao, Y. and Karypis, G. (2005), "Hierarchical clustering algorithms for document datasets", *Data Mining and Knowledge Discovery*, Vol. 10 No. 2, pp. 141-68.

Zhu, M.J., Shi, S.M., Li, M.J. and Wen, J.R. (2009), "Effective top-*k* computation with term-proximity support", *Information Processing & Management*, Vol. 45 No. 4, pp. 401-12.

Further reading

Gordon, M.D. and Lenk, P. (1991), "A utility theoretic examination of the probability ranking principle in information retrieval", *Journal of the American Society for Information Science*, Vol. 42 No. 10, pp. 703-14.

Hua, M., Pei, J., Fu, A.W.C., Lin, X. and Leung, H. (2009), "Top-*k* typicality queries and efficient query answering methods on large databases", *The VLDB Journal*, Vol. 18 No. 3, pp. 809-35.

Huang, A. (2008), "Similarity measures for text document clustering", *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 2008*, Vol. 2008, pp. 49-56.

Korenius, T., Laurikkala, J. and Juhola, M. (2007), "On principal component analysis, cosine and Euclidean measures in information retrieval", *Information Sciences*, Vol. 177 No. 22, pp. 4893-905.

Robertson, S.E. (1977), "The probability ranking principle in IR", *Journal of Documentation*, Vol. 33 No. 4, pp. 294-304.

Tang, X.H., Chen, G.Q. and Wei, Q. (2009), "Introducing relation compactness for generating a flexible size of search results in fuzzy queries", *Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, Lisbon, Portugal*, Vol. 2009, pp. 1462-7.

About the authors

Baojun Ma is a PhD student in the Department of Management Sciences and Engineering, School of Economics and Management, Tsinghua University, Beijing, China. His research interests focus on information retrieval techniques, business intelligence and data mining. Baojun Ma is the corresponding author and can be contacted at: mabj.03@sem.tsinghua.edu.cn

Qiang Wei is an Associate Professor in the Department of Management Sciences and Engineering, School of Economics and Management, Tsinghua University, Beijing, China. His teaching and research interests include knowledge discovery, business intelligence and data mining, management information systems, soft computing, e-business and online advertising.

Guoqing Chen is China's National Chang-Jiang Scholars' Professor in Information Systems at the Department of Management Sciences and Engineering, School of Economics and Management, Tsinghua University, Beijing, China. His teaching and research interests include business intelligence and decision support, e-business, and soft computing techniques.