A COMBINED MEASURE FOR REPRESENTATIVENESS ON INFORMATION RETRIEVAL IN WEB SEARCH*

BAOJUN MA† QIANG WEI GUOQING CHEN

School of Economics and Management, Tsinghua University, Beijing, 100084, China

Information retrieval is one of the important technologies adopted by search engines. This paper discusses the representativeness of a small set of search results extracted from the original results, which is deemed desirable in web search, and then proposes a combined measure, namely RF_β to evaluate the extracted small set in terms of the notions of coverage and redundancy. Data experimental results on Google, TCW and Random extraction strategies show that the RF_β measure could effectively evaluate the representativeness considering users' preferences.

1. Introduction

Information retrieval (IR) [1] is one of the important technologies adopted by search engines [2]. Since different users may have discrepant search requirements even with same keywords, different criteria are used for evaluating the quality of IR. Many quality measures are concerned with the match between keywords and search results [3], such as recall [4], precision [4] and some combined measures (e.g. F-value [5], R-precision [6], MAP [6], NDCG [7]).

In real-world applications, a search engine (e.g., a keyword search) may often generate a huge volume of records or pages, all of which could hardly be browsed one-by-one by users. Usually, users would prefer a small set of search records that appear in the first couple of web pages and have a good quality of the search. This is deemed meaningful and important to Internet 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 records to appear (or

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

† Corresponding author, Email: mabj.03@sem.tsinghua.edu.cn

to appear first, if many) as a web search result is of great interest to both academia and practitioners. In this paper, we focus on the quality of search, in light of representativeness, in two respects: one is the coverage rate referring to the number of the elements in the original set "covered" by the extracted small sets; the other is the redundancy rate referring to the number of "redundant" elements in the extracted small set. These two rates reflect the representativeness of the small set on the original set in two different ways. Merely for illustrative purposes, let us first consider a simplified case where the original collection of records is {A, B, C, C, D}, then two small collections {A, B, C, C} and {A, B, C} refer to the same number of records (i.e., 4 out of 5), whereas the former is more redundant than the latter. In usual search cases, the records (or documents/texts etc.) are generally close to each other (e.g., via text similarity). It is easy to encounter a situation where some records are almost identical or highly close. 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 texts, each being a cited report on the same news from a single media source, which reads similar but hardly interesting. As a matter of fact, a combined view of representativeness in coverage and redundancy is deemed desirable, which motivates the effort of this study.

2. Evaluation measures of representativeness

Let us concentrate our discussion on web search for documents. Given a set of n web documents, e.g., $D = \{d_1, d_2, \ldots, d_n\}$, where d_i is a web document, $i = 1, \ldots, n$, then an IR method E is to extract m documents from D ($m \leq n$) 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 m, a good IR method is to extract a D^E with as large coverage rate as possible and as small redundancy rate as possible.

2.1. *Coverage rate*

For two documents 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 closeness, which is reflexive and symmetric. For example, a commonly used measure in IR is the keyword-based Cosine similarity, i.e., $F_c(d, d') = \cos\theta = \langle d, d' \rangle / (|d||d'|)$, where $\langle d, d' \rangle$ is a dot product of d and d' and $|d|$ is a magnitude of d [9]. 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:

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

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^E_1 = {A, B, C, C} and D^E_2 = {A, B, C}. Then we have $r_C(D^E_1, D) = 4/5$, and $r_C(D^E_2, D) = 4/5$, because both cover 4 documents in D. Furthermore, in a closeness case with the closeness measure $F_C(d', d) = |d' \cap d|/4$ (i.e., $|d' \cap d|$, which represents the number of elements that d' and d share), then suppose D = {ABCD, ABCE, FGHI, FGHI, FGHJ}, D^E_1 = {ABCD, ABCE, FGHJ} and D^E_2 = {ABCE, FGHI}, we have $r_C(D^E_1, D) = 9/10$, and $r_C(D^E_2, D) = 9/10$ as well. Note that both D^E_1 and D^E_2 cover the same number of documents. However, they have different levels of redundancy, which will be discussed in the next section.

Moreover, the coverage rate has the following properties:

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

where D, D' and D'' are nonempty sets of documents.

2.2. *Redundancy rate*

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

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

Referring to the above examples, in the crisp case, $r_R(D^E_1) = 1/4$ and $r_R(D^E_2) = 0$. In the closeness case, $r_R(D^E_1) = 2/7$ and $r_R(D^E_2) = 0$. In a combined view, in either case, D^E_2 is considered better in representativeness than D^E_1 since D^E_2 has the same degree of coverage as D^E_1 but a lower degree of redundancy than D^E_1. Moreover, for any nonempty set D, the redundancy rate has the following properties:

- $0 \le r_R(D) < 1$.
- $r_R(D) = 0$, if $|D| = 1$.

- $r_R(D) = 0$, if $F_C(d, d') = 0$, $\forall d, d' \in D$, $d \neq d'$, $|D| > 1$.
- $r_R(D) = 1 - 1/|D|$, if $F_C(d, d') = 1$, $\forall d, d' \in D$, $d \neq d'$.
- If a document d, $d \in D$, is with $1-1/\sum_{d' \in D} F_C(d', d) > r_R(D)$, then $r_R(D - \{d\}) < r_R(D)$.

2.3. *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 [4-5], 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 $RF_\beta(D^E_1) = 24/31 < 8/9 = RF_\beta(D^E_2)$, while in the closeness case, $RF_\beta(D^E_1) = 90/113 < 18/19 = RF_\beta(D^E_2)$, which conforms to the fact that D^E_2 has a higher level of representativeness that D^E_1.

3. Data experiments

In order to verify the RF_β measure, data experiments were conducted to compare the search results of Google search engine and a representative IR method proposed in [8], namely TCW, as well as a Random extraction 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., m documents ($m << 1000$), to search their preferred documents. Though Google's first m documents were with high PageRank values and might have been diversified considering similarity,

many documents were still found quite similar, e.g., the hottest content relevant to keywords usually appears frequently in different documents, which may imply high information coverage but high redundancy. The TCW method is to cluster *D* into *m* information-equivalent classes and extract one representative document for each class, which tries to obtain low redundancy without significant loss of information coverage [8]. Furthermore, a random extraction strategy (hereafter called Random) is to randomly extract *m* documents in *D* with uniform distribution. Moreover, the Cosine similarity measure in the vector space IR model [9-10] is used to obtain the degree of closeness between web documents.

In the experiments, several keywords are randomly chosen. The values of the coverage rates and redundancy rates are shown in Table 1 with m = 10, 20, and 30, respectively, which approximately represent 1, 2, and 3 web pages. For Random extraction, in order to narrow the deviations, the listed values are the means of 50 IID extractions. Note that we conducted the experiments on a 3.00GHz 2.96Gb 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, http parser and http client packages and APIs.

Table 1 The coverage rates and redundancy rates of search results

		***m* = 10**			***m* = 20**			***m* = 30**		
Keyword	**Rate**	**Google**	**TCW**	**Random**	**Google**	**TCW**	**Random**	**Google**	**TCW**	**Random**
argument	d_C	**0.1947**	0.1252	0.1786	**0.2418**	0.1504	0.2444	0.2769	0.1732	**0.2898**
	d_R	0.4413	**0.1614**	0.2989	0.5385	**0.2273**	0.4715	0.5686	**0.4098**	0.5733
capital	d_C	**0.1704**	0.1500	0.1681	0.2125	0.1709	**0.2193**	0.2438	0.1906	**0.2547**
	d_R	0.4209	**0.1480**	0.3223	0.4763	**0.3129**	0.4860	0.5391	**0.4299**	0.5806
logistic	d_C	**0.2419**	0.1552	0.2139	0.2788	0.1839	**0.2853**	0.3107	0.2086	**0.3264**
	d_R	0.4675	**0.1983**	0.3525	0.6130	**0.4000**	0.5428	0.7231	**0.5094**	0.6390
operation	d_C	0.1301	0.1197	**0.1590**	0.1654	0.1377	**0.2072**	0.2295	0.1561	**0.2470**
	d_R	0.3357	**0.0655**	0.3095	0.4490	**0.2197**	0.4483	0.5128	**0.3371**	0.5562
origin	d_C	0.1111	0.1239	**0.1737**	0.1951	0.1486	**0.2358**	0.2318	0.1804	**0.2802**
	d_R	0.2595	**0.1284**	0.2973	0.4647	**0.2868**	0.4630	0.5418	**0.3930**	0.5747
party	d_C	0.1394	0.3178	**0.3362**	0.2079	0.3447	**0.4106**	0.2409	0.3686	**0.4596**
	d_R	0.3791	**0.1282**	0.3302	0.5749	**0.2727**	0.5291	0.6369	**0.4348**	0.6224

* The bold value represents the highest d_C or lowest d_R among Google, TCW and Random.

Table 1 shows that, first, the redundancy rates of TCW were always the lowest, meaning that TCW could cope with redundancy most effectively, but its coverage rates tended to be the lowest. Second, in most cases, Random had good performance on coverage superior to Google and TCW, especially when *m* was large. Further, with the increase of *m*, both of these two rates of the three

strategies would keep increase, since the larger the subset is, the higher possibility to cover information and generate redundancy.

Additionally, by setting β = 0 (100% preference to coverage), 0.5, 1 (equal preference to coverage and non-redundancy), 2, 10 and 100 (almost 100% preference to non-redundancy), more experiments were conducted to further examine the correspondingly RF_β values (m = 10, 20, 30). Consistently with Table 1, the experiments revealed that if users preferred high coverage (i.e., β < 1), Google and Random showed better representativeness than TCW, and that if users preferred high non-redundancy (i.e., β > 2), TCW performed better than Google and Random on representativeness. Therefore, as discussed previously, the proposed 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.

4. Conclusion

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. Data experiments were conducted to compare three IR strategies, namely Google, TCW and Random, showing their different performances. Future studies could center on constructing an IR method for extracting representative information based on the RF_β measure.

References

1. Bing Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* (Springer Berlin Heidelberg, New York, 1998).
2. A. Spink and B.J. Jansen, *Web search: public searching of the web* (Kluwer Academic Publishers, 2004).
3. M.J. Zhu, S.M Shi, M.J Li and J.R. Wen, Effective top-k computation with term-proximity support, *Information Processing & Management* 45, 401(2009).
4. D. E. Kraft and A. Bookstein, Evaluation of Information Retrieval System: A Decision Theory Approach, *Journal of the American Society for Information Science* 29: 31–40 (1978).
5. C.J van Rijsbergen, *Information Retrieval (2nd Edition)* (Butterworths, London, 1979).
6. C. Buckley and E.M. Voorhees, Evaluating evaluation measure stability, in *Proceedings of the 23rd ACM SIGIR conference*, 2000.
7. T. Sakai, On the reliability of information retrieval metrics based on graded relevance, *Information Processing & Management* 43(2): 531–548 (2007).
8. X.H. Tang, G.Q. Chen, Q. Wei, Introducing Relation Compactness for Generating a Flexible Size of Search Results in Fuzzy Queries, in *Proceedings of the Joint IFSA and EUSFLAT conference*, 2009.
9. G. Salton, *The SMART retrieval system: Experiments in automatic document processing*, Englewood Cliffs, N. J. (Prentice-Hall, 1971).
10. C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval* (Cambridge University Press, 2009).