An Evaluation of New and Old Similarity Ranking Algorithms

Paul Lynch, Xiaocheng Luan, Maureen Prettyman, Lee Mericle, Edward Borkmann, and Jonathan Schlaifer
U. S. National Library of Medicine, Lister Hill National Center for Biomedical Communications, Computer Science Branch
plynch@mail.nih.gov, luan@nlm.nih.gov, reenie@lhc.nlm.nih.gov

Abstract

The National Library of Medicine's (NLM) IRVIS project has been evaluating "similarity ranking" algorithms that re-order search results according to their similarity to a target result. Several variations of known ranking algorithms were tested, as well as one (we believe) new one which weights terms based on word length. When the algorithms were evaluated using the OHSUMED test collection, the new word length based algorithm was found to outperform the others.

1. Introduction

The Information Retrieval and Visualization (IRVIS) project is exploring new ways of organizing and presenting search results, characterizing available data, and assisting the user in formulating queries. The research's primary focus is in the context of search systems that provide access to more than one source or type of information, such as the NLM Gateway (http://gateway.nlm.nih.gov). As a part of that work, we needed an algorithm for measuring the similarity of one search result's text to another. Because we are developing modules which existing systems could use, the data available for the ranking is limited to a sub-set of the available search results (e.g., the first 50 results returned), and the algorithm needs to be able to perform the ranking on the fly. The following nine algorithms are the subject of the testing being reported here.

2. Algorithms

Each of the rankers (ranking algorithms) takes as input the ***target*** text string (the search result for which similar results are desired), and list of ***context*** strings (the subset of search results provided by the client system). The rankers work by comparing each context string against the target, getting a similarity score for each, and then ranking the results according to the scores.

2.1. Set Ranker

This ranker computes a similarity score by treating the two strings as sets of (unique) terms **A** and **B**, and taking as the score **S** the ratio of the number of terms in the sets' intersection to the number of terms in their union.

2.2. Word Length

This ranker is very much like the Set Ranker, except that the terms are weighted according to their length, i.e.:

$$S \equiv \frac{\sum_{x \in A \cap B} length(x)}{\sum_{x \in A \cup B} length(x)}$$

This algorithm is based on the hypothesis that, in general, longer words are more likely to represent the subject of a text string than are shorter words.

2.3. Aslam-Frost

J. Aslam and M. Frost [1] have proposed an information-theoretic approach to measuring the similarity between strings of text based on work by D. Lin [2]. A ranker was created based on their formula.

2.4. Simple Vector

Vector-space ranking algorithms were introduced by Salton [3]. This is the simplest such algorithm, in which terms were completely unweighted.

2.5. Vector SW,N,IDF,PML

This is 2.4, but with four types of term weights are applied (in sequence): Stop Words, Normalization, Inverse Document Frequency, and PubMed's local term weight. These four weights are described below.

Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)
0-7695-2108-8/04 $ 20.00 © 2004 IEEE

2.5.1. Stop Words (SW). Each element of the vector that corresponds to a term found in a set of 366 commonly occurring words is set to zero.

2.5.2. Normalization (N). Each number in the string's vector is divided by the vector-space length of the vector.

2.5.3. Inverse Document Frequency (IDF). Each element in the vector is multiplied by a factor which is smaller if the term for that element occurs in many of the text strings in the context. Specifically, the formula given in [4] was used.

2.5.4. PubMed Local (PML). This is a local term weight used by the PubMed website that gives more weight to a term that occurs frequently in a particular text string [5].

2.6. Vector SW,IDF,PML

This is 2.5, without normalization.

2.7. Vector SW,N,IDF,TF

This is the same as 2.5 except for the substitution of the Term Frequency (TF) local term weight. The form used was the augmented normalized term frequency [6].

2.8. Vector SW,IDF,TF

This is 2.7, without normalization.

2.9. Vector SW

This is 2.4, with the Stop Words term weight applied.

3. Testing the Algorithms

We decided to test the similarity ranking algorithms using the OHSUMED test collection [7] (developed by Dr. William Hersh, and others) from Oregon Health Sciences University. The collection consists of 348,566 MEDLINE records (journal article citations), 106 queries, the retrieved ***documents*** (records) for those queries, and relevance judgments for the retrieved documents. In our testing of the algorithms, we made the assumption that documents judged relevant to the same query were also likely to be relevant to each other.

For a given query and ranking algorithm, we ranked the query's retrieval set and gave the ranking a score based on how close the ranking was to an "ideal" ranking based on the relevance judgments. By using each query and trying different documents as target strings for the algorithms, we obtained 2,126 scores for each algorithm.

4. Results

Table 1 shows the average score of the algorithms across all 2126 trials. Statistical tests showed that even though numerically the scores are close, the top three scores are significantly different from each other and from the others. Trials were done with and without using the records' abstracts; in both cases the Word Length algorithm outperformed the others. For more details, see http://irvis.nlm.nih.gov.

Table 1. Summary of Results

		Average Score	Std. Dev.
1	**Word Length**	0.69216	0.10742
2	**Set Ranker**	0.67992	0.10596
3	**Aslam-Frost**	0.65967	0.10544
4	**Vector SW,N,IDF,TF**	0.61925	0.10215
5	**Vector SW,N,IDF,PML**	0.61731	0.10207
6	**Simple Vector**	0.61725	0.10542
7	**Vector SW**	0.61717	0.10776
8	**Vector SW,IDF,TF**	0.61536	0.10094
9	**Vector SW,IDF,PML**	0.61385	0.10096

5. References

[1] J. A. Aslam and M. Frost, "An Information-theoretic Measure for Document Similarity", *26th ACM Annual SIGIR Proceedings*, pages 449-450, 2003.

[2] D. Lin. An Information-Theoretic Definition of Similarity. In Proceedings of International Conference on Machine Learning, 1998.

[3] G. Salton and M. E. Lesk, "Computer Evaluation of Indexing and Text Processing", *Journal of the ACM*, 15(1):8-36, 1968.

[4] W. John Wilbur and Yiming Yang, "An Analysis of Statistical Term Strength and its Use in the Indexing and Retrieval of Molecular Biology Texts", *Comput. Biol. Med.*, Vol. 26, No. 3, 1996, p. 210.

[5]http://www.ncbi.nlm.nih.gov/entrez/query/static/computation.html, 6/4/2003.

[6] Wilbur & Yang, p. 211

[7] W. Hersh, C. Buckley, T. J. Leone, D. Hickam. "OHSUMED: an interactive retrieval evaluation and new large test collection for research", *17th Annual ACM SIGIR Proceedings*, pages 192-201, 1994.