CM-PMI: Improved Web-based Association Measure with Contextual Label Matching

Xiaojun Wan
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
wanxiaojun@icst.pku.edu.cn

ABSTRACT
WebPMI is a popular web-based association measure to evaluate the semantic similarity between two queries (i.e. words or entities) by leveraging search results returned by search engines. This paper proposes a novel measure named CM-PMI to evaluate query similarity at a finer granularity than WebPMI, under the assumption that a query is usually associated with more than one aspect and two queries are deemed semantically related if their associated aspect sets are highly consistent with each other. CM-PMI first extracts contextual labels from search results to represent the aspects of a query, and then uses the optimal matching method to assess the consistency between the aspects of two queries. Experimental results on the benchmark Miller Charles' dataset demonstrate the good effectiveness of the proposed CM-PMI measure. Moreover, we further fuse WebPMI and CM-PMI to obtain improved results.

Categories and Subject Descriptors:
H.4.m [**Information Systems**]: Miscellaneous

General Terms: Algorithms, Experimentation

Keywords: Association measure, Web mining, CM-PMI

1. METHOD
The study of measuring semantic similarity between words or entities has become very important for many web-related tasks, including word clustering, query reformulation and substitution, name disambiguation, community mining, etc. In recent years, web-based association measures have been well studied to evaluate the semantic similarity between two words or entities. In contrast with knowledge-based measures relying on existing knowledge databases or taxonomies (e.g. WordNet), web-based measures make use of the up-to-date web search results returned by web search engines and they can reflect the updated semantic similarity between two words or entities. Moreover, web-based measures can be successfully applied to compute the semantic similarity between new words or entities, which are usually not defined in any existing knowledge database.

A number of web-based similarity measures have been proposed in recent years, including WebPMI [1, 6], CODC [2], web-based kernel [5] and supervised learning [1]. WebPMI is the most popular one used today and it uses the number of hits returned by a web search engine for assessing the semantic similarity between two queries q_1 and q_2 as follows:

$$WebPMI(q_1,q_2)=\begin{cases}0 & \text{if } \text{hits}(q_1 \text{ and } q_2)\le c\\ \log_2 \dfrac{hits(q_1 \text{ and } q_2)/N}{(hits(q_1)/N)\cdot(hits(q_2)/N)} & \text{otherwise}\end{cases} \quad (1)$$

where N is the number of web pages indexed by the search engine and c is a threshold. In this study we set N=10^{10} and c=5 as in [1].

WWW 2008, April 21--25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

hits() returns the hits page count for the query, estimated by the search engine.

The above WebPMI measure directly computes the similarity between two queries. However, we observe that for each short query, the number of the returned search results is usually very large and the results usually contain diverse aspects about the query. For example, some results for the query "bike" are about bike photos, while other results are about bike components. Therefore, the search results for a query can be organized into a few subtopics about the query, each subtopic representing a specific aspect of the query. We believe that the subtopics in the search results for a query can reflect the query at a fine-grained level, while the single topic representation of the search results for a query in previous work is coarse-grained. The more the subtopic sets of two queries are consistent with each other, the more the queries are semantically similar with each other. In this study, we use a contextual label of a query to represent a subtopic for the query. A contextual label is actually a word which occurs frequently nearby the query in the search results. Figure 1 gives the contextual labels for synonyms "bike" and "bicycle", and the two words share many common aspects linked by the dash line.

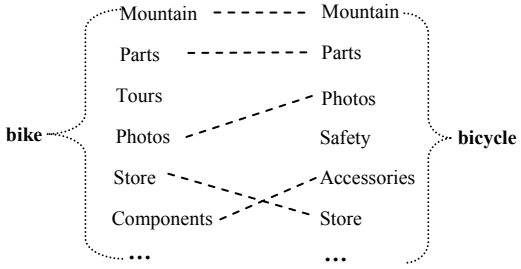

Figure 1. Contextual label matching example

Based on the above assumption, we propose a novel association measure named CM-PMI to make full use of the aspect information of each query. The measure evaluates semantic similarity between words or entities in an indirect way. It first extracts the contextual labels in the search results for each query and then measures the consistency between the contextual label sets. The optimal matching method is employed to measure the consistency between the contextual label sets by formalizing the problem as the optimal matching problem in the graph theory. The normalized optimal matching weight is used as the semantic similarity between the queries.

The CM-PMI measure consists of three steps: search results retrieval, contextual label extraction and contextual label matching. They are respectively described as follows:

Search Results Retrieval
In this study, we base our experiments on Microsoft's Live Search (http://search.live.com), without loss of generality. Live Search is one of the most popular search engines used today and it returns

the estimated hits number and at most 1000 results for each query. We extract each result record consisting of the title, snippet and url from the returned result page. Instead of downloading the full web pages, we use only the titles and snippets in the search results for efficient computation. All the 1000 (or less) returned search results are used in this study.

Contextual Label Extraction

The contextual label words are extracted for each query in the following simple way: the stopwords are removed from the titles and snippets of all the search results, and the remaining words that co-occurs with the query are ranked in decreasing order of their TF.IDF scores, and finally the top m ($m \geq 1$) words are chosen as the contextual labels of the query. The size of the co-occurrence window is typically set to 2 words. Each label is deemed to reflect one aspect of the query.

Contextual Label Matching

Given two sets of contextual labels X and Y for two queries q_1 and q_2, this step aims to measure the consistency between the label sets from a global perspective. We formalize this problem as the optimal matching problem and allow only one-to-one matching between the contextual labels. A globally optimal solution can be achieved by solving the optimal matching problem.

Optimal matching (OM) is a classical problem in the graph theory. Let $G=\{X, Y, E\}$ be a weighted bipartite graph, where $X=\{x_1,x_2,...,x_m\}$ and $Y=\{y_1,y_2,...y_m\}$ are the partitions representing the two sets of contextual labels for queries q_1 and q_2. $V=X \cup Y$ is the vertex set, and $E=\{e_{ij} | 1 \leq i,j \leq m\}$ is the edge set with each edge e_{ij} connects a contextual label x_i in X and a contextual label y_j in Y. A weight w_{ij} is assigned to every edge e_{ij} in G. The weight w_{ij} is computed using the WebPMI measure as follows:

$$w_{ij} = WebPMI(x_i, y_j) \quad (2)$$

A matching M of G is a subset of the edges with the property that no two edges of M share the same node. Given the weighted bipartite graph G, OM is to find the matching $\tilde{M}$ that has the largest total weight. The Kuhn-Munkres algorithm [3] is employed to solve the OM problem. Lastly the optimal matching $\tilde{M}$ in graph G is acquired and we use the normalized total weight in $\tilde{M}$ as the semantic similarity value between queries q_1 and q_2:

$$CM-PMI(q_1,q_2) = \frac{\sum_{e_{ij} \in \tilde{M}} w_{ij}}{\min\{|X|,|Y|\}} \quad (3)$$

where $\min\{|X|,|Y|\}$ returns the minimum size of X and Y. Here, we have $\min\{|X|,|Y|\}=m$.

2. EVALUATION

In the experiments, we further propose the FusionPMI meaure to fuse the WebPMI and CM-PMI scores as follows:

$$FusionPMI(q_1,q_2) = \lambda \cdot WebPMI(q_1,q_2) + (1-\lambda) \cdot CM-PMI(q_1,q_2) \quad (4)$$

where $\lambda \in [0,1]$ is the fusion weight. We compare CM-PMI, WebPMI and FusionPMI based on the Miller-Charles dataset [4], which contains 30 word-pairs rated by a group of 38 human subjects. The word pairs are rated on a scale from 0 (no similarity) to 4 (perfect synonymy). We follow previous researches and use only 28 pairs for evaluation because two word pairs are omitted in earlier versions of WordNet. We use the Pearson's correlation coefficient (r) to measure the correlation between automatic computed values and human-labeled values.

Figure 2 shows the comparison results with respect to various label number (i.e. $m \in [2,20]$). We can see that CM-PMI outperforms WebPMI when a medium-size number of label words (i.e. $m \in [8,18]$) are used to reflect the aspects of a query. Very small or very large label number will deteriorate the performance of CM-PMI, which is because that too few labels cannot cover all the important aspects of a query and too many labels can introduce noisy aspects. FusionPMI can almost always outperform both WebPMI and CM-PMI, which demonstrates that WebPMI and CM-PMI can complement each other. Figure 3 shows the performance curves for the FusionPMI measures (m=5,10,15) with respect to the fusion weight λ. We can see that FusionPMI can always outperform WebPMI when m is appropriately set (e.g. 10 or 15) and CM-PMI plays a more important role than WebPMI for FusionPMI. Overall, the results demonstrate the good effectiveness of both the CM-PMI measure and the FusionPMI measure.

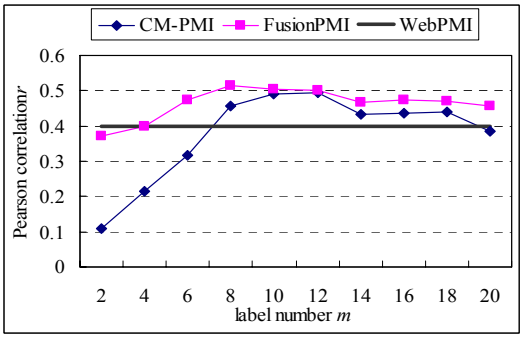

Figure 2. Pearson correlation (r) vs. label number (m)

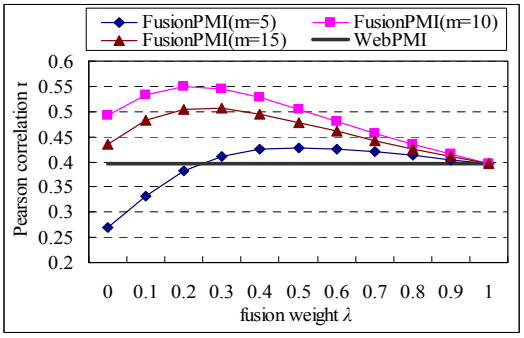

Figure 3. Pearson correlation (r) vs. fusion weight (λ) for FusionPMI

3. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China (No.60703064) and the Research Fund for the Doctoral Program of Higher Education of China (No.20070001059).

4. REFERENCES

[1] D. Bollegala, Y. Matsuo and M. Ishizuka. Measuring semantic similarity between words using web search engines. In Proceedings of WWW2007.

[2] H.-H. Chen, M.-S. Lin and Y.-C. Wei. Novel association measures using web search with double checking. In Proceedings of COLING-ACL2006.

[3] H. W. Kuhn. The Hungarian method for the assignment problem, Naval Res. Logist. Quart. 2: 83–97, 1955.

[4] G. Miller and W. Charles. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1):1-28, 1998.

[5] M. Sahami and T. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In Proc. of WWW 2006.

[6] P. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of ECML2001.