Journal of Southeast University (English Edition) Vol. 22, No. 3, pp. 389 – 393 Sept. 2006 ISSN 1003—7985

Ontology-based similarity measure for text clustering

Yan Duanwu[1] Li Xiaopeng[2] Wang Lei[1] Cheng Xiao[1]

([1] Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China)

([2] Library, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: A method that combines category-based and keyword-based concepts for a better information retrieval system is introduced. To improve document clustering, a document similarity measure based on cosine vector and keywords frequency in documents is proposed, but also with an input ontology. The ontology is domain specific and includes a list of keywords organized by degree of importance to the categories of the ontology, and by means of semantic knowledge, the ontology can improve the effects of document similarity measure and feedback of information retrieval systems. Two approaches to evaluating the performance of this similarity measure and the comparison with standard cosine vector similarity measure are also described.

Key words: similarity measure; text clustering; ontology; information retrieval system

Information retrieval systems (IRS) such as web search engines and document database retrieval systems are playing an important role in people's daily lives. However, most current IRS, especially web search engines are mainly based on keyword-based search/retrieval and may retrieve documents which are not relevant to a query, or may not retrieve all relevant documents.

To address this issue, some search engines also include category-based searches. Obviously, a search engine will be more user friendly and easier to maintain if there is a system that automatically generates and uses human-like categories. Much research related to the issue has been done and Frakes et al. [1-2] developed the classic references. In the most recent IRSs, documents are modeled using the vector space, where the product of term frequency (TF) and the inverse document frequency (IDF) have been proposed as term-weighting schemes. The cosine vector similarity and Euclidean distance measures are widely applied. A clustering is a type of classification of a set of objects. There are hierarchical and partitioned clustering methods. Some implementations of clustering algorithms were proposed in Refs. [3 – 4]. Clustering has been applied to documents in Refs. [5 – 6] for organization, summarization and location of topics. Other artificial intelligence approaches to classification of documents have been reported in Refs. [7 – 8] where machine learning is emphasized.

WordNet[9] and HowNet[10] are of lexical database, where words are organized by different types of relationships. Such thesauri can be used as an input ontology to our clustering technique. In Refs. [11 – 12], related work has been done, where words are clustered to produce such relationships. This paper focuses on the use of ontology-based similarity measures that combine category and keyword-based information to generate document clusters of improved quality.

1 Similarity and Distance Measures for Documents

1.1 Cosine vector similarity measure

In the case of a t-dimensional space, if the number of distinct keywords in the total collection of document is t, and a document d_i in the collection is represented by the vector $\boldsymbol{d}_i = \{w_{i1}, w_{i2}, \ldots, w_{it}\}$, then the cosine vector similarity between document d_i and d_j is defined as

$$\mathrm{sim}(d_i, d_j) = \frac{\sum_{k=1}^{t} (w_{ik} w_{jk})}{\sqrt{\sum_{k=1}^{t} w_{ik}^2} \sqrt{\sum_{k=1}^{t} w_{jk}^2}} \tag{1}$$

Baeza-Yates et al. [2] gave a motivation for the cosine vector similarity measure by arguing that the sum of the products of the frequency of corresponding keywords (or vector dot product) between two documents is a good indicator of the similarity between those documents. It was also pointed out that, to avoid the problem of discriminating against smaller documents in favor of larger ones, a normalization by the length of the document vector is in order. Now, we first set out to

Received 2006-04-25.

Foundation item: The Young Teachers Scientific Research Foundation (YTSRF) of Nanjing University of Science and Technology in the Year of 2005—2006.

Biography: Yan Duanwu (1976—), male, doctor, lecturer, yanwu_nju@163.com.

understand how normalization affects document similarity measure.

Definition 1 Let d_i be any document represented by the vector $\boldsymbol{d}_i = \{w_{i1}, w_{i2}, \ldots, w_{it}\}$, and let a be any positive real number. We define the scaling normalization of document d_i by a (scalar-document product) as the document $\frac{1}{a}d_i$, represented by the vector $\frac{1}{a}\boldsymbol{d}_i$. The k-th component of the normalized vector is

$$w'_{ik} = \frac{w_{ik}}{a} \tag{2}$$

Proposition 1 The cosine vector similarity measure is invariant with respect to scaling.

Proof Let d_i and d_j be any two documents represented by vectors $\boldsymbol{d}_i = \{w_{i1}, w_{i2}, \ldots, w_{it}\}$ and $\boldsymbol{d}_j = \{w_{j1}, w_{j2}, \ldots, w_{jt}\}$ in the t-dimensional space, and let a and b be any two positive real numbers. We shall prove that

$$\mathrm{sim}\left(\frac{d_i}{a}, \frac{d_j}{b}\right) = \mathrm{sim}(d_i, d_j) \tag{3}$$

where $\mathrm{sim}(d_i, d_j)$ is the cosine vector similarity between documents d_i and d_j.

$$\mathrm{sim}\left(\frac{d_i}{a}, \frac{d_j}{b}\right) = \frac{\sum_{k=1}^{t}\left(\frac{w_{ik}}{a}\frac{w_{jk}}{b}\right)}{\sqrt{\sum_{k=1}^{t}\left(\frac{w_{ik}}{a}\right)^2}\sqrt{\sum_{k=1}^{t}\left(\frac{w_{jk}}{b}\right)^2}} =$$

$$\frac{\frac{1}{ab}\sum_{k=1}^{t} w_{ik}w_{jk}}{\sqrt{\frac{1}{a^2}\sum_{k=1}^{t} w_{ik}^2}\sqrt{\frac{1}{b^2}\sum_{k=1}^{t} w_{jk}^2}} = \mathrm{sim}(d_i, d_j) \tag{4}$$

This result is another form of the property of normalized cross correlation being independent of scale factors in Ref. [4].

1.2 Euclidean distance measure and normalization by document length

Considering that there are t different keywords in the collection of documents, if the following vector in the t-dimensional space represents a document d_i: $\boldsymbol{d}_i = \{w_{i1}, w_{i2}, \ldots, w_{it}\}$, then the Euclidean distance can define the dissimilarity between two documents d_i and d_j as

$$\mathrm{dis}(d_i, d_j) = \sqrt{\sum_{k=1}^{t}(w_{ik} - w_{jk})^2} \tag{5}$$

From Eq. (5), we can see that the Euclidean distance for the domain of documents shows that the documents with comparable size tend to be more similar to each other and less similar to other much larger or much smaller documents, even though all these documents are semantically similar. In the following, we try to address this problem by normalizing the documents vectors by their vector length.

Let document d_i be represented by the t-dimensional vector $\boldsymbol{d}_i = \{w_{i1}, w_{i2}, \ldots, w_{it}\}$. We define the length of document d_i as the length of vector $\boldsymbol{d}_i$, that is

$$|d_i| = |\boldsymbol{d}_i| = \sqrt{\sum_{k=1}^{t} w_{ik}^2} \tag{6}$$

We define the normalized document vector as

$$\boldsymbol{d}'_i = \frac{1}{|\boldsymbol{d}_i|}\boldsymbol{d}_i \tag{7}$$

The k-th component of the normalized vector is then

$$w'_{ik} = \frac{w_{ik}}{\boldsymbol{d}_i} \qquad k = 1, 2, \ldots, t \tag{8}$$

The Euclidean distance between the document vectors normalized by their length can define the modified dissimilarity between two documents d_i and d_j as follows:

$$\mathrm{dis}'(d_i, d_j) = \sqrt{\sum_{k=1}^{t}(w'_{ik} - w'_{jk})^2} \tag{9}$$

Proposition 2 The Euclidean distance applied to the normalized document vectors is semantically equivalent to the standard cosine vector similarity measure. In fact, the cosine vector similarity is quadratically proportional to the Euclidean distance.

Proof Let d_i and d_j be any two documents represented by vectors $\boldsymbol{d}_i = \{w_{i1}, w_{i2}, \ldots, w_{it}\}$ and $\boldsymbol{d}_j = \{w_{j1}, w_{j2}, \ldots, w_{jt}\}$ in the t-dimensional space. We shall prove that

$$\mathrm{sim}(d_i, d_j) = 1 - \frac{\mathrm{dis}'^2(d_i, d_j)}{2} \tag{10}$$

where $\mathrm{sim}(d_i, d_j)$ is the cosine vector similarity between documents d_i and d_j, and $\mathrm{dis}'(d_i, d_j)$ is the Euclidean distance between the normalized document vectors of $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$.

$$\mathrm{dis}'^2(d_i, d_j) = \sum_{k=1}^{t}(w'_{ik} - w'_{jk})^2 =$$

$$\sum_{k=1}^{t}\left(\frac{w_{ik}}{\sqrt{\sum_{l=1}^{t} w_{il}^2}} - \frac{w_{jk}}{\sqrt{\sum_{l=1}^{t} w_{jl}^2}}\right)^2 =$$

$$\sum_{k=1}^{t}\left(\frac{w_{ik}^2}{\sum_{l=1}^{t} w_{il}^2} - \frac{2w_{ik}w_{jk}}{\sqrt{\sum_{l=1}^{t} w_{il}^2}\sqrt{\sum_{l=1}^{t} w_{jl}^2}} + \frac{w_{jk}^2}{\sum_{l=1}^{t} w_{jl}^2}\right) =$$

$$1 - 2\frac{\sum_{k=1}^{t}(w_{ik}w_{jk})}{\sqrt{\sum_{k=1}^{t} w_{ik}^2}\sqrt{\sum_{k=1}^{t} w_{jk}^2}} + 1 = 2 - 2\mathrm{sim}(d_i, d_j) \tag{11}$$

This equation is another form of the cosine law in Ref. [4].

1.3 Discussion

According to the analysis above, we have shown that the Euclidean distance applied to the normalized document vector is semantically equivalent to the standard cosine vector similarity measure. While the Euclidean distance has the property of being invariant with respect to rotation and translation, the cosine vector measure has the property of being invariant with respect to scaling. In the domain of documents, the latter property is more important since, for example, we expect a technical paper or a news report to have a high similarity measure with regard to their abstract or shorter versions. Moreover, with the advance in the development of ontologies and domain specific topical hierarchies, as in most recent web search engines, the ability to automatically use the semantics provided by such knowledge structures is desirable. For those reasons, we propose a hybrid similarity measure that combines both keyword-based information and ontology-based knowledge.

2 Ontology-Based Similarity Measure

As discussed in the previous section, the use of ontology-based knowledge is motivated by the essence of and the intrinsic properties of the cosine vector similarity measure itself. Keywords that are part of the same category or topic should mutually contribute with a greater effect on the similarity of two documents in which they appear. In this section, we define the proposed ontology-based similarity measure (OBSM) and a general description of the two approaches used to evaluate the performance of OBSM that has also been provided.

2.1 Principles of the similarity measure

An ontology in the text space can be defined as a hierarchy of categories. Each category is defined by a set of keywords, ordered by the degree of importance with respect to their category. This order is achieved by associating a weight that measures the importance of a keyword in that category. Keywords may be part of many categories. This allows for an overlapping hierarchical category structure. The ontology relationships are obtained by considering the terms as part of concepts. Combining the ontology with standard keyword-based similarity, a new semantic to the information retrieval system is added. The ontology can be constructed from a term categorization structure such as a thesaurus, which generally includes synonyms. However, relationships between keywords in the ontology may be general. Ideally, a domain specific expert should specify them.

If t is the number of distinct keywords in the whole collection of documents, s is the number of categories in the ontology, w_{ik} is the term weight associated with keyword k and document d_i, and w'_{ck} is the weight of keyword $k(k=1,2,\ldots,t)$ with respect to a category c, then the ontology-based similarity between two documents d_i and d_j represented by the t-dimensional vectors $\boldsymbol{d}_i=\{w_{i1}, w_{i2}, \ldots, w_{it}\}$ and $\boldsymbol{d}_j=\{w_{j1}, w_{j2}, \ldots, w_{jt}\}$ is

$$\mathrm{sim}(d_i, d_j) = \frac{\sum_{c=1}^{s}\left(\sum_{k=1}^{t} w_{ick}\sum_{k=1}^{t} w_{jck}\right)}{\sqrt{\sum_{c=1}^{s}\left(\sum_{k=1}^{t} w_{ick}\right)^2\sum_{c=1}^{s}\left(\sum_{k=1}^{t} w_{jck}\right)^2}} \quad (12)$$

where $w_{ick}=w_{ik}w'_{ck}$.

The original weight associated with keyword k and document d_i is multiplied by the relative weight of keyword k with respect to category c. The sum of the products is performed over all categories. Different keywords can be identified if they are in the same category. This guarantees that if there exists a category to which many common keywords of documents d_i and d_j belong, then the contribution of those keywords is expanded, causing the two documents to be more similar to each other than to other documents. Therefore, those documents will have a better chance of being classified in a cluster that semantically resembles the category of the input ontology. It is worth noting that we keep the normalization by the vector length, so that again we avoid discrimination against smaller documents in favor of larger ones.

The OBSM can also be used for the similarity between documents and queries in the case of interactive information retrieval systems, like a search engine. In that case, the user may even contribute interactively in the choice of the categories that may be used to expand the contribution of the keywords that are entered in the query.

2.2 Performance evaluation approaches

We considered two different evaluation approaches. In the first approach, we generate clustered data (here, clustered is in terms of keywords overlap). This is similar to the training set approach used in artificial intelligence based clustering. Then we compute the distance between these clusters in terms of similarity measures, using standard cosine vector similarity measures (CVSM) and OBSM, and we check if the clusters are preserved. In the second approach, we collect real documents and cluster them using both similarity measures.

In the first approach, to evaluate the performance

of our technique using OBSM, we need to generate a set of documents, with keywords drawn from a limited set of keywords. An ontology is built associating keywords into different categories, with weights representing an ordering within a category. A document is then associated with a category by having more keywords drawn from that category than any other. We vary the overlap between documents in terms of the percentage of common keywords, both within the same cluster (intra-cluster overlap) and from different clusters (inter-cluster overlap). If a cluster i has n_i t-dimensional points $\{\boldsymbol{d}_{ik}\}$ $(k = 1, 2, \ldots, n_i)$, then the intra-cluster and inter-cluster average similarities are defined respectively as follows:

$$d_{ii} = \frac{1}{n_i(n_i - 1)} \sum_{k=1}^{n_i} \sum_{l=1, l \neq k}^{n_i} \mathrm{sim}(d_{ik}, d_{il}) \quad (13)$$

$$d_{ij} = \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \mathrm{sim}(d_{ik}, d_{jl}) \quad (14)$$

For various data sets, we can compute the intra-cluster and inter-cluster average document similarities. A data set with a proper clustered structure should exhibit a high intra-cluster similarity, showing a proper compactness of the clusters, and a low inter-cluster similarity, showing a proper isolation between clusters.

The second approach is completely different from the previous one. In fact, in this approach, we need to collect real documents with no *a priori* knowledge in terms of initial cluster structure. Then we cluster the documents employing a clustering technique, using both similarity measures: CVSM and OBSM. This is different from the first approach, where we do not perform a clustering. Instead, we need to generate clustered data, where the clustered structure is measured by overlap in terms of common keywords. Then, we compute the dissimilarity between these clusters in terms of both similarity measure (standard cosine vector and the proposed similarity measure), and we check if and how well the cluster structure is preserved.

3 Conclusion

Ontology is a valid method to represent domain specific concepts and their semantic relations under hierarchical categories, and can bring semantic knowledge into information retrieval systems among information resource organizing, to perform improved similarity measures for document clustering. The Euclidean distance applied to the normalized document vector is semantically equivalent to the standard cosine vector similarity measure. Motivated by the essence of and intrinsic properties of the cosine vector similarity measure, we propose a hybrid similarity measure that combines both keyword-based information and ontology-based knowledge and also provide a general description of the two approaches that can be used to evaluate the performance of OBSM.

References

[1] Frakes W B, Baeza-Yates R. *Information retrieval data structure and algorithms* [M]. Englewood Cliffs, New Jersey: Prentice Hall, 1992.

[2] Baeza-Yates R, Ribeiro-Neto B. *Modern information retrieval* [M]. Translated by Wang Zhijin. Beijing: China Machine Press, 2005. (in Chinese)

[3] Fisher D. Iterative optimization and simplification of hierarchical clusterings [J]. *Journal of Artificial Intelligence Research*, 1996, **27**(4): 147 − 179.

[4] Frigui H, Nasraoui O. Simultaneous clustering and dynamic keyword weighting for text documents [A]. In: Berry M W, ed. *Survey of Text Mining* [C]. New York: Springer, 2003. 45 − 72.

[5] Klose A, Nurnberger A, Kruse R, et al. *Interactive text retrieval based on document similarities*, *physics and chemistry of the earth*, *part A*: *solid earth and geodesy* [M]. Amsterdam, Netherlands: Elsevier, 2000. 649 − 654.

[6] Uzuner O, Davis R, Katz B. Using empirical methods for evaluating expression and content similarity [EB/OL]. (2004-12-30) [2006-02-10]. http://people. csail. mit. edu/ozlem/Uzuner-HICSS04. pdf.

[7] Shyu M L, Chen S C, Shu C M. Affinity-based probabilistic reasoning and document clustering on the WWW [A]. In: *Proceedings of the* 24*th IEEE Computer Society International Computer Software and Applications Conference* [C]. Washington, DC: IEEE Computer Society, 2000. 149 − 154.

[8] Yaniv R, Souroujon O. Iterative double clustering for unsupervised and semi-supervised learning [EB/OL]. (2002-12-30) [2006-02-20]. http://books. nips. cc/papers/files/nips14/AA24. pdf.

[9] Niles I, Pease A. Linking lexicons and ontologies: mapping wordNet to the suggested upper merged ontology [A]. In: *Proceedings of the International Conference on Information and Knowledge Engineering* [C]. Las Vegas, Nevada, 2003. 161 − 172.

[10] Gan K W, Wong P W. Annotating information structures in Chinese text using HowNet [A]. In: *Proc of the* 2*nd Chinese Language Processing Workshop*, *Association for Computational Linguistics Conference* [C]. Hong Kong, 2002. 85 − 92.

[11] Wulfekuhler M R, Punch W. Finding salient features for personal Web page categories [EB/OL]. (2001-07-23) [2005-12-20]. http://www. cps. msu. edu/wulfekuh/research/PAPER118. ps.

[12] Slonim N, Tishby N. Document clustering using word clusters via the information bottleneck [A]. In: *Proc of the*

23*rd Annual Intl ACM SIGIR Conf on Research and Development in Information Retrieval*[C]. Athens, Greece, 2000. 208 - 215.

文本聚类中基于本体的相似性测度

颜端武[1] 李晓鹏[2] 王 磊[1] 成 晓[1]

([1] 南京理工大学信息管理系,南京 210094)
([2] 南京理工大学图书馆,南京 210094)

摘要:介绍了一种综合各层级分类类目和对应关键词来构造概念体系并用于改进信息检索系统效果的方法. 为了改进文本聚类的效果,提出了将领域知识本体和文本关键词词频相结合的基于余弦向量的文本相似性测度方法. 该本体面向特定领域,将关键词以不同权值对应于各分类类目,通过其语义知识来改进文本相似性测度以及信息检索系统的效果. 进一步给出了对基于本体的相似性测度方法进行效果评价的 2 种策略以及该方法与经典余弦向量测度方法的比较结果.

关键词:相似性测度;文本聚类;本体;信息检索系统

中图分类号:TP391. 1