A Statistics-Based Semantic Relation Analysis Approach for Document Clustering

Xin Cheng, Duoqian Miao, and Lei Wang

Department of Computer Science and Technology, Tongji University, Shanghai, China
cx1227@gmail.com, maioduoqian@163.com

Abstract. Document clustering is a widely research topic in the area of machine learning. A number of approaches have been proposed to represent and cluster documents. One of the recent trends in document clustering research is to incorporate the semantic information into document representation. In this paper, we introduce a novel technique for capturing the robust and reliable semantic information from term-term co-occurrence statistics. Firstly, we propose a novel method to evaluate the explicit semantic relation between terms from their co-occurrence information. Then the underlying semantic relation between terms is also captured by their interaction with other terms. Lastly, these two complementary semantic relations are integrated together to capture the complete semantic information from the original documents. Experimental results show that clustering performance improves significantly by enriching document representation with the semantic information.

1 Introduction

Document clustering aims to organize the documents into groups according to their similarity. The traditional approaches are mostly based on Bags of words (BOW) model, which represents the documents with the terms and their frequency in the document. However, this model has the limitation that it assumes the terms in the document are independent thus regardless of the semantic relationship between them. It considers the documents are dissimilar if no overlapped terms exist, even though they describe the same topic.

To overcome the disadvantage of BOW model, a lot of approaches have been proposed to capture the semantic relation between terms to enhance document clustering. Generally, there are two directions to explore the semantic relation between terms: knowledge-based approach and statistics-based approach [3][6][7][13]. The knowledge-based approach measures the semantic relation between terms using the background knowledge which is constructed from ontology, such as WordNet [12] and Wikipedia [6]. Although the incorporation of the background information into BOW model has shown an improvement in document clustering, this approach has the limitation that the coverage of the ontology is limited, even for WordNet or Wikipedia. Besides, the context information has been overlooked to compute the semantic relation between terms. The statistics-based approach captures the semantic relation between terms based on term co-occurrence information, which evaluates the semantic relation between terms from the significance of their co-occurrence pattern. The most previous statistics-based

D. Miao et al. (Eds.): RSKT 2014, LNAI 8818, pp. 332–342, 2014.
DOI: 10.1007/978-3-319-11740-9_31

approaches only capture the explicit semantic relation between terms from their co-occurrence information, but the underlying relation has been overlooked, which is also essential for capturing the complete semantic relation between terms. Besides, the synonymous and ambiguous terms could not be accurately handled in the previous approaches, and that would affect the accuracy of semantic relation evaluation in a certain degree.

In this paper, we propose a novel approach to capture the semantic relation between terms based on both the explicit and implicit relations between terms. It firstly captures the explicit relation between terms from their co-occurrence information, and then the implicit semantic relation is revealed by their interaction with other terms. Meanwhile, Wikipedia is exploited to handle the synonymous and ambiguous terms. Lastly, the explicit and implicit semantic relations are integrated to capture the complete semantic information from the original documents, and then we extends the original BOW model with the semantic information for document clustering.

The rest of the paper is organized as follows. Section 2 presents the background of document clustering problem and reviews some related work. Section 3 proposes a novel approach for mining the semantic relation between terms and analyzing the semantic information of the original documents. The experimental results are discussed in Section 4, and the conclusion and future work will be describe in Section 5.

2 Related Work

Document clustering is an unsupervised approach to group the similar document together, and most document clustering approaches are based on the BOW (Bag of Words) model, which assumes that the terms in the document are independent. However, the terms are always related to each other, and the related information between them could be hierarchical relationship, compound word relation and synonym relation etc.

The semantic relation between terms was first introduced by Wong for document representation [14], and then many approaches are proposed to measure the relation between terms. Some approaches have been proposed to explore the semantic relation between terms with background knowledge, like WordNet and Wikipedia. In [2], they proposed to measure the relatedness between terms not by the exact term matching, but by their semantic relation, which is measured based on the semantic information in WordNet. However, WordNet has the limited coverage because it is manually built. In [7], Wikipedia, the largest electronic encyclopaedia, was exploited for document clustering. They construct a proper semantic matrix based on the semantic relation between terms from the underlying structural information in Wikipedia, and then they incorporated the semantic matrix into traditional document similarity measure.

Another direction of term relation measure is based on the statistical information. Examples of such work like the generalized vector space model (GVSM), which was proposed by Wong et al. [14], captures the semantic relation between terms in an explicit way by using their co-occurrence information. It simply utilizes the document-term matrix W^T as the semantic matrix S, and then each document vector is projected as $d^{'} = d * W^T$. The corresponding kernel between two document vectors is expressed as $k^{'}(d_i, d_j) = d_i W^T W d_j$. The entry in matrix $W^T W$ reflects the similarity between terms which is measured by their frequency of co-occurrence across the document collection, which means two terms are similar if they frequently co-occur in the same document. Holger et al. [1] uses term co-occurrence patterns to estimate term dependency. It integrates the semantic information into document representation for calculation of the document similarity. The empirical results confirm that it improves the performance of document retrieval for particular document collections. Argyris et al. [8] take the local distance of the co-occurrence terms into consideration while computing the relation between terms. They exploit the relation between terms in the local context, and then combined all the local relation together to constitute the global relation matrix.

3 Methodology

The BOW model exploits each term in document as document features, so it cannot model efficiently the rich semantic information of documents. To capture the accurate similarity between documents, its essential to build a high quality document representation which could reserve the semantic information from the original documents. A lot of work have proposed that if two terms co-occur in the same document, they are relational in a certain degree [5][8][11]. However, they just consider the explicit relation of terms in the same document, but the underlying relation between them has been overlooked, which is also essential to capture the robust and reliable relation between terms. In our approach, a novel approach is proposed to capture the relation between terms, which identifies the relation between terms by not only themselves, but also their interaction with other terms.

In our work, we propose a novel semantic analysis model. This model capitalizes on both the explicit relation and implicit relation to compute the semantic relation between terms. The key points of the proposed model are: (a) it computes the semantic relation between each pair of terms using their co-occurrence information as the explicit relation; (b) it further constructs semantic links between terms by considering their interaction with other terms as the implicit relation; and (c) it combines the explicit and implicit relations together to compute the semantic relation for each pair of terms. Using this model, the semantic relation between terms can be captured more precisely, which can be integrated into document representation to enhance the quality of document representation.

3.1 The Semantic Relation Analysis between Terms

The first step of our approach for measuring the semantic relation between terms is to explore the explicit semantic relation. In most of the previous approaches, the relation between terms is simply estimated by considering the co-occurrence frequency but overlooking the discriminative power of terms, which will lead to the incorrect estimation of the relation between terms. In this work, the $tfidf$ scheme is used to measure the relation between terms which is based not only on the frequency of terms but also on their discriminative ability. Firstly, we introduce the definition of the explicit relation between terms:

Definition 1. *Let D be a document collection, two terms t_i, t_j are considered to be explicitly related only if they co-occur in the same document. To evaluate the explicit relation between two terms, we propose an efficient measure which is defined as:*

$$Relation_{exp}(t_i, t_j) = \frac{1}{|H|} \sum_{d_x \in H} w_{xi} w_{xj} / (w_{xi} + w_{xj} - w_{xi} w_{xj}) \tag{1}$$

Where w_{xi} and w_{xj} are the $tfidf$ values of term t_i, t_j in the document d_x, and H denotes the documents where t_i and t_j co-occur.

With the explicit relation between terms, the quality of document representation can be enhanced by integrating the explicit relation into document representation. However, the underlying relation between terms cannot be discovered from term co-occurrence information. In the following, we will introduce a novel approach to capture the implicit relation between terms:

Definition 2. *Let D be a document collection, two terms t_i, t_j are from different documents ($t_i \in d_m, t_j \in d_n$), if there is a term t_s co-occur with them in the respective documents, they are considered as being linked by term t_s, and they are implicitly related.*

Fig. 1 shows an example of term implicit relation analysis, two terms t_i and t_j are from different document, and t_{s1},t_{s2} are the co-occurrence terms with them in the respective documents. Terms t_i and t_j are not related based on the explicit relation analysis, but they are considered to be relational using the implicit relation analysis because they co-occur with the same terms in the respective documents. Therefore, we define the calculation of the implicit relation between terms as follows:

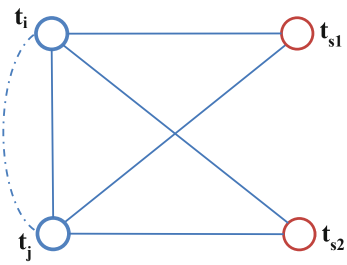

Fig. 1. An example of the implicit relation analysis

Definition 3. *Let D be a document set, a pair of terms (t_i, t_j) are from different documents, the relation between t_i and t_j can be linked by t_s which is the same co-occurrence terms with t_i and t_j in the respective documents. The implicit relation between t_i and t_j, by their interaction with their co-occurrence term $t_s \in S$, is defined as:*

$$Relation_{imp}(t_i, t_j) = \frac{1}{|S|} \sum_{t_s \in S} \frac{\min((Relation_{exp}(t_i, t_s), Relation_{exp}(t_j, t_s))}{\sum_{t_x \in T}(Relation_{exp}(t_x, t_s)}, \tag{2}$$

Where $Relation_{exp}(t_i, t_s)$, $Relation_{exp}(t_j, t_s)$ represent the explicit relation of the term t_i and t_j with term t_s in the respective documents, and S is the term collection which t_i and t_j co-occur with, T is the term collection of this corpus.

Term Sense Disambiguation. It is essential to measure whether an ambiguous term takes the same sense in different documents. That is because if two terms co-occur with an ambiguous term and it takes different sense in each document, then they could not be considered that they co-occur with the same term, which means the co-occurrence term could not be taken as the link term.

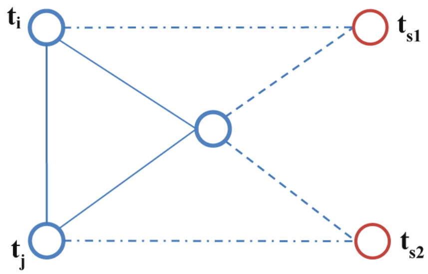

Fig. 2. An example of the relation with equivalent terms

As Fig. 2 demonstrates, term t_i and t_j co-occurrence with the same term t_s, but t_s takes different sense t_{s1} and t_{s2} in the respective documents, so t_i and t_j could not be linked by the term t_s.

To alleviate this problem, we explore the intersection of their surrounding text to disambiguate the sense of terms, because the context information is an indication of the sense of each term, and the terms with the same sense should appear in the similar contexts. The sense similarity can be evaluated by two main steps: context information extraction and similarity evaluation. We first identify the context information from the co-occurrence matrix, as all the co-occurrence terms with each term is considered to be the context information. Then the similarity of the sense is defined as:

$$sim(s_1, s_2) = (|N(s_1) \cap N(s_2))|)/(|N(s_1)| + |N(s_2)|) \tag{3}$$

Where $N(s_i)$ represents all the co-occurrence terms with term s_i, and $N(s_1) \cap N(s_2)$ is the common co-occurrence terms between s_1 and s_2. In our approach, if $sim(s_1, s_2) < 0.5$, term t_s is considered as an ambiguous term, which means terms t_i and t_j can not be linked by t_s.

Mapping of Equivalent Terms. In some cases, two terms are similar even same in sense but differs in spelling. For example, “disk” and “disc”, “motor” and “engine”, “BBC” and “British Broadcasting Corporation”, and they should be taken as the same term because they are just the alternative names, alternative spellings or abbreviations of the same thing.

Like in Fig. 3, t_i co-occurs with t_{s1} while t_j co-occurs with the term t_{s2}, t_{s1} and t_{s2} are not same in appearance, like "Car" and "Automobile", but they have the same meaning of term t_s, then it is intuitive that t_i and t_j should be considered as being related as they co-occur with the same term t_s.

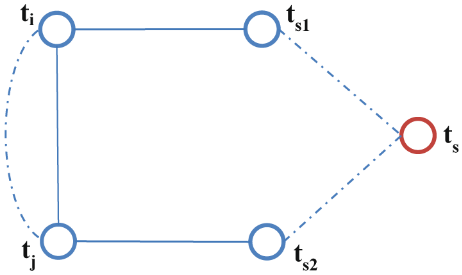

Fig. 3. An example of the relation with polysemous words

To solve this problem, its essential to map the equivalent terms to the identical expression. In our paper, we take Wikipedia, which has been proved to be an efficient thesaurus, as background knowledge to solve this problem. In Wikipedia, the redirect hyperlinks group the terms that have the same sense together and link to the identical concept, and they are very useful as an additional source of synonyms. Hence, if two terms link to the indexical concept, they are considered as being the link term between t_i and t_j.

The explicit relation discovers the relation between terms by using their co-occurrence statistics and the implicit relation discovers the relation between terms by using their interaction with other terms. To capture the complete semantic relation between terms, we integrate the explicit and implicit relations together to measure the semantic relation between terms in this section.

Definition 4. *Let D be a document collection, terms t_i and t_j appear in this document collection, then the semantic relation between t_i and t_j is defined as:*

$$Relation(t_i, t_j) = Relation_{exp}(t_i, t_j) \cdot Relation_{imp}(t_i, t_j)), \tag{4}$$

where $Relation_{exp}(t_i, t_j)$ is explicit relation between t_i and t_j, and $Relation_{imp}(t_i, t_j)$ is the implicit relation between t_i and t_j.

In our approach, the co-occurrence statistics are modeled with the integration of explicit and implicit relations. In this sense, our approach has the advantage of capturing the complete semantic relation between terms from term co-occurrence statistics. Furthermore, the semantic relation matrix can be constructed which reflects the semantic relation between each pair of terms, and then it can be used to project the original document representation into a new feature space with better discriminative ability.

3.2 The Document Semantic Analysis

Based on the proposed semantic relation analysis, the semantic matrix S can be further constructed whose elements reflect the semantic relation between each pair of terms.

With the semantic matrix S, the original documents can be mapped into a new feature space, which reserves the semantic information from the original documents.

$$d : \boldsymbol{d} \mapsto \boldsymbol{d}' = \boldsymbol{d} * S, \tag{5}$$

By integrating the semantic information into document representation, the original documents can be mapped into a new feature space. In the new feature space, the documents are well distinguished and it can further improve the performance of the related document analysis task.

4 Experiment and Evaluation

In this section, we empirically evaluate our approach with document clustering, and the BOW is used as the baseline for comparison. To focus our investigation on the representation rather than the clustering method, we used the standard k-means algorithm in the experiments.

4.1 Data Sets

To validate our strategy, we conduct experiments on four document collections. D1 is the subset of 20 Newsgroups while D2 is the mini-newsgroup version, D3 is the subsets of Reuters 21578, and D4 is the WebKB document collection. The detailed information of these document collections is described as follows:

Table 1. Characteristics of Data Sets

Data sets	Name	Classes	m	n	n_{avg}
D1	20 newsgroup	5	1864	16516	76
D2	20 newsgroup	20	1989	24809	55
D3	Reuters21578	8	2091	8674	33
D4	WebKB	4	4087	7769	32

1. The first data set (D1) is a subset of 20 Newsgroups(20NG), which is a widely used data set for document clustering [9]. It consists 1864 newsgroup documents across 5 classes.
2. The second data set (D2) is the mini-newsgroups version, which has $1,989$ documents across all 20 classes in 20-newsgroups.
3. The third data set (D3) is a subset derived from the popular Reuters-21578 document collection [10] which has $2,091$ documents belonging to 8 classes (acq, crude, earn, grain, interest, money-fx, ship, trade).
4. The last data set (D4) is WebKB [4]. It consists of 4087 web pages and manually classified into 4 categories.

4.2 Evaluation Criteria

Cluster quality is evaluated by four criterions: purity, rand index, F1-measure and normalized mutual information.

Purity is a simple and transparent way to measure the quality of clustering. The purity of a cluster is computed by the ratio between the size of the dominant class in the cluster and the size of cluster. $purity(c_i) = \frac{1}{|c_i|} \max_j |c_j|$.Then the overall purity can be expressed as the weighted sum of all individual cluster purity:

$$purity = \frac{|c_i|}{N} \sum_{i=1}^{n} purity(c_i), \tag{6}$$

Rand Index (RI) measures the clustering quality by the percentage of the true positive and true negative decisions in all decisions during clustering:

$$RI = ((TP + TR))/((TP + TR + FP + FR)) \tag{7}$$

where TP (true positive) denotes that two similar documents are assigned to the same cluster; TN (true negative) denotes that two dissimilar documents are assigned to different clusters; FP (false positive) denotes that two dissimilar documents are assigned to the same cluster, and FN (false negative) denotes that two similar documents are assigned to different clusters.

F1-measure considers both the precision and recall for clustering evaluation:

$$F1 = ((precision * recall))/((precision + recall)) \tag{8}$$

where $precision = TP/(TP + FP)$, $recall = TP/(TP + FN)$.

Normalized mutual information (NMI) is a popular information theoretic criterion for evaluating clustering quality. It is computed by dividing the Mutual Information between the entropy of the clusters and the label of dataset:

$$NMI(C, L) = (I(C; L))/(H(C) + H(L))/2) \tag{9}$$

where C is a random variable for cluster assignments, L is a random variable for the pre-existing classes on the same data. $I(C; L)$ is the mutual information between the clusters and the label of the dataset, and $H(C)$ and $H(L)$ is the entropy of C and L.

4.3 Performance Evaluation

Table 2 shows the performance of our proposed approach on each dataset compared with two other approaches: the classic BOW model and the GVSM model, and the classic BOW model is taken as the baseline for comparison. For these quality measures, higher value in [0, 1] indicates better clustering solution. We can observe that our approach achieves significant improvement in all quality measures. Compared with the base line, our proposed approach has achieved 10.4%, 22.7%, 11.1% and 19.4% average improvement. Compared to GVSM model, our approach also achieves 7.4%, 16.9%, 8.8% and 15.5% average improvement. The experimental results demonstrate the benefit of integrating both the explicit and implicit probabilistic relation between

Table 2. Document Clustering Results by Using K-means

Data Sets	Purity			RI			F1-measure			NMI		
	BOW	*GVSM*	*CRM*	*BOW*	*GVSM*	*CRM*	*BOW*	*GVSM*	*CRM*	*BOW*	*GVSM*	*CRM*
D1	0.293	0.325	**0.413**	0.340	0.461	**0.541**	0.356	0.351	**0.417**	0.139	0.158	**0.403**
D2	0.125	0.114	**0.189**	0.447	0.447	**0.760**	0.123	0.122	**0.197**	0.207	0.198	**0.325**
D3	0.740	0.775	**0.821**	0.669	0.691	**0.817**	0.594	0.567	**0.749**	0.421	0.447	**0.597**
D4	0.431	0.495	**0.581**	0.357	0.448	**0.604**	0.455	0.478	**0.505**	0.094	0.216	**0.312**

terms into document representation. Although the GVSM model is assisted by the proposed semantic smoothing, which takes into account the local contextual information associated with term occurrence, it overlooks the underlying semantic relation between terms. Compared to the GVSM model, our proposed approach considers both the explicit and implicit relations between terms, which can capture more reliable semantic relation between terms.

An interesting point to stress according to Table 2 is that larger gains are obtained in the document collections which are harder to classify, where the baseline does not perform well. For example, for the D1 and D2 collections, which are more difficult to obtain good clustering results using only bag-of-words representation. By integrating the semantic information captured with our approach into document representation, the clustering results have been significantly improved.

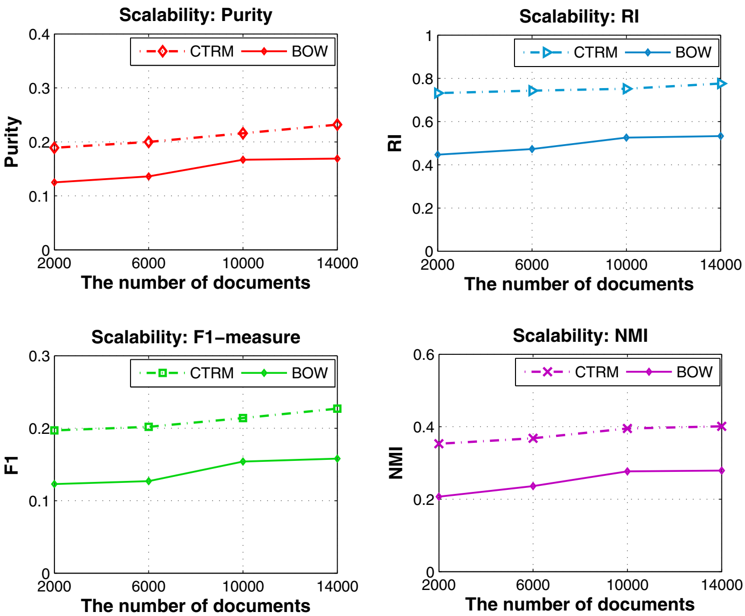

Fig. 4. The impact of corpus size

Besides, even in the cases where the performance of baseline is good and improvements consequently tend to be more limited, we also achieve statistically significant gains. Likewise, for D3, we still achieves 8.1%, 14.8%, 15.5% and 17.6% gains.

4.4 The Impact of Corpus Size

In this subsection, we analyze the effect of corpus size on the semantic relation analysis of our approach. To show the effect of corpus size, we conduct a set of experiments on the document collection 20-newsgroups by increasing the number of documents from $2,000$ to $14,000$ at increments of $4,000$.

The experimental results are shown in Fig. 4. It is interesting to note that our approach achieves significant gains compared to the baseline on the small collection with 2, 000 documents. Meanwhile, with the increase in the document collection size, the performance of our approach shows a slightly higher improvement over the baseline. In summary, the experimental results show that our strategy augments performance on different sizes of document collection, even on the small document collection, and the improved performance is stable with the increasing size of document collection.

5 Conclusion and Future Work

This paper presents a novel approach for the semantic relation analysis. In this approach, the semantic relation between terms is measure based on both the explicit and implicit relations. The experiment results indicate that our approach can significantly improve the performance of document clustering.

In the future, we will work on three aspects to improve our approach: (1) the independence test is essential to determine whether two terms co-occur together more often than by chance; (2) the optimal integration of the explicit and implicit relations can be further improved; (3) the reduction of time complexity is worthy further analysis.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (No. 61075056, 61273304), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20130072130004) and the Fundamental Research Funds for the Central Universities.

References

1. Billhardt, H., Borrajo, D., Maojo, V.: A context vector model for information retrieval. Journal of the American Society for Information Science and Technology 53(3), 236–249 (2002)
2. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics 32(1), 13–47 (2006)
3. Bullinaria, J.A., Levy, J.P.: Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods 39(3), 510–526 (2007)
4. Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Slattery, S.: Learning to extract symbolic knowledge from the world wide web. In: Proceedings of the 15th National Conference on Artificial Intelligence (1998)

5. Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M.A., Meira Jr, W.: Word co-occurrence features for text classification. Information Systems 36(5), 843–858 (2011)
6. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI, vol. 7, pp. 1606–1611 (2007)
7. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 389–396 (2009)
8. Kalogeratos, A., Likas, A.: Text document clustering using global term context vectors. Knowledge and Information Systems 31(3), 455–474 (2012)
9. Lang, K.: Newsweeder: Learning to filter netnews. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 170–178 (1995)
10. Lewis, D.D.: Reuters-21578 text categorization test collection, distribution 1.0 (1997), `http://www.research.att.com/~lewis/reuters21578.html`
11. Burgess, C., Lund, K.: Modelling parsing constraints with high-dimensional context space. Language and cognitive processes 12(2-3), 177–210 (1997)
12. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
13. Wang, P., Hu, J., Zeng, H.J., Chen, Z.: Using wikipedia knowledge to improve text classification. Knowledge and Information Systems 19(3), 265–281 (2009)
14. Wong, S.K.M., Ziarko, W., Wong, P.: Generalized vector spaces model in information retrieval. In: SIGIR 1985. pp. 18–25. ACM (1985)