Effect of Pronoun Resolution on Document Similarity

Atul Kumar
Indian Institute of Information Technology
Allahabad India

Sudip Sanyal
Indian Institute of Information Technology
Allahabad India

ABSTRACT

This paper presents a novel effect of Pronoun Resolution on measurement of document similarity. In this paper we have studied the effect of pronoun resolution within the framework of the Vector Space Model and Probabilistic Latent Semantic Analysis. For this purpose we have developed a Benchmark Corpus consisting of documents whose similarity scores have been given by human beings. We measured the inter-document similarity on these documents using VSM and PLSA. We then performed pronoun resolution on these documents and again calculated the similarity using both methods. Next, the correlation coefficient of the scores was taken with those of the human generated scores. The correlation coefficients clearly demonstrated substantial and consistent improvements of the similarity score after pronoun resolution.

Categories and Subject Descriptors

D.3.3 [Natural Language Processing]: Document Similarity

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Document Similarity, Pronoun Resolution, Information Retrieval, Statistical Algorithm

1. INTRODUCTION

The world is accessing the information available in millions of TB data. Today it is one of the great challenges in information sciences to develop intelligent interfaces for human beings which support computer users in their quest for relevant information. Given the above real world scenario, it is important to be able to measure the similarity among a pair of documents in order to retrieve the relevant documents from a given collection of documents. The above limitations are particularly true for unstructured text databases. The conventional information retrieval techniques often yield results that are not truly relevant. Thus, it is pertinent to experiment with Natural Language Processing in order to find whether the retrieval results can be improved. In this paper we propose a novel technique of Information Retrieval that incorporates Natural Language Processing techniques like Pronoun Resolution. Measurement of document similarity has been used in various areas like natural language processing, text categorization, information retrieval and information extraction.

It also finds application in essay grading [9], relevance feedback [17] and finding similar articles in electronic newspaper with the help of cross lingual dictionary.

Various models and algorithm have been proposed in the past decade for measuring the similarity score between documents. Various algorithms are available to compute similarities between documents [19]. One of the earliest approaches to Document similarity is perhaps the vector model, where the document most similar to an input document (presented in the form a query document) is determined by ranking the documents in a different type collection documents in the reverse order of their similarity to the given input document [1, 3, 4]. Some researchers used word-to-word document similarity with help of either knowledge bases [15, 16] or corpus base [14, 18]. For document based semantic similarity, conceivably the most widely used approaches are the approximations through query expansion or the latent semantic analysis [5,6] that measures the similarity of documents by exploiting second order word relations automatically acquired from the corpus.

A related line of work consists of approaches, computing document similarity score using phrase indexing graph model [18] that have made need of the tree edit distance method to compute similarity. Wan and Peng apply the earth mover's distance calculates to compute document similarity [22]. Some researcher recently used Fuzzy measure for document similarity measure [21]. Fuzzy document similarity system utilizes fuzzy sets to represent document membership function degrees for query term similar, fuzzy logical operators to specify queries and fuzzy compatibility calculates to evaluate the retrieval status value of a document. Probabilistic Latent Semantic Analysis (PLSA) attempts to find hidden concepts in the corpus [7,8] and it seems to be the most promising till date. It uses the expectation maximization method [10] to extract the latent concepts in a document using purely statistical methods. Theses latent concepts can then be used to find a score between a pair of documents.

Our primary focus in this work is to study the effect of Pronoun Resolution (PR) on the similarity score of documents. To see why we expect PR to play an important role, consider the following case. Let there be an article on Nehru. It is quite likely that after the first sentence we refer to Nehru as “he”. Most of the existing methods would remove “he” as a stop word in the preprocessing stage. Even if it is not removed, the term “he” would get a high score and the term “Nehru” would get a low score. This would affect the precision because the word “he” might have been used in a different context in another document. It would also affect recall because documents with a small count for “Nehru” (but with a high count of “he” which refer to “Nehru”) would not be retrieved even though they are quite relevant.

Based on this observation, in this paper we study the effect of Pronoun Resolution (PR) on document similarity. We use the framework of PLSA and VSM for performing our tests.

This paper is organized as follows in different sections. In section 2 we discuss the VSM and PLSA. In section 3 we walk through a small example (a traditional story), calculating document similarity score in documents before PR and after PR using VSM and PLSA. In section 4 we discuss the main results of this paper where we first describe the specific experiments that we perform. We then analyze the results by calculating the Spearman correlation coefficient of the document similarity scores obtained using the automatic methods with the similarity scores given by humans (the ground truth). The final section summarizes our main conclusions and suggests directions for further experimentation.

2. SIMILARITY SCORING TECHNIQUE

As mentioned in the section 1, we would like to study the effect of PR on the document similarity score. In order to make certain that the change is due to PR and not an artifact of the specific algorithm used for calculating the similarity; we calculate the similarity using two different techniques namely the Vector Space Model and the Probabilistic Semantic Analysis technique. Thus, in the present section we briefly describe these two techniques.

2.1 Vector Space Model

In 1975 Gerald Salton [3] [4] proposed a statistical, "Vector Space Model", which works by representing the documents in an 'n' dimensional space, where 'n' is number of different terms or words (as $t_{1,}$ $t_{2,}$ t_3 .... t_n) which consists of the whole vocabulary of the corpra or collection of document. Each document is considered as a vector D_1, $D_{2.....}$ D_c is the 'n' dimensional space. Here 'c' is total number of document in the corpora. Document Vector can be shown as following:

$$D_r = \{d_{1r}, d_{2r}, d_{3r}, \ldots\ldots d_{nr}\}$$

Where d_{ir} is the i^{th} factor of the vector representing the r^{th} document [2]. The Vector Space Model is mostly used in such a case where the documents collection are placed in the term space and VSM is required to find the similar document for a given query document. A query is similar to a small document. The similarity among the query document and collection of documents is calculated and the best suitable matching documents are returned. Various similarity measures have been proposed in past decade. The one that is very frequently used is cosine similarity where the cosine similarity the query document vector, Q, and a vector of document 'D' is calculated as:

$$\mathrm{Cos}\ \Theta = Q*D/|Q|*|D|$$

In the traditional Vector Space Model, the tf*idf process is used to find out the weight of the term in specified document vector i.e. the components of the vector, d_{ir}. It basically depends on two main factors.

1. The frequency of occurrence of term 'i' in the document 'r' (term frequency $tf_{r,i}$)

2. The frequency of occurrence of term 'i' in the document collection (document frequency df_i).

So, the weight of a term 'i' in given document 'r' can be written as

$$W_{r,i} = tf_{r,i} * idf_i = tf_{r,i} * \log(c/df_i)$$

Where

c = Number of documents in the collection of document

Idf = Inverse document frequency

Tf = term frequency

This model integrates both the local and global information of entire system. The first term, $tf_{r,i}$ accounts for local weight while the ratio (df_i /c) is the probability of selecting a document that contains a queried term from the documents-collection. The ratio can be treated as global probability for the whole collection.

2.2 Probabilistic Latent Semantic Analysis

Thomas Hofmann has given a statistical model for indexing technique in 1999[7][8] called the Probabilistic Latent Semantic Analysis. The core of PLSA is the aspect model. The aspect model is hidden variable model for co-occurrence data which associates hidden class variables b€ B = { b_1 , b_2 , b_3,……..} for each observation, i.e. with each occurrences of following terms t € T = { t_1,t_2 ,….} in a particular context document d € D= {d_1,d_2…..}. The probabilities related to this model are defined as follows.

P (d) denotes a probability of selecting a document d in given documents.

P (t|b) denotes a probability of generating a term t in hidden class b.

P (b|d) denotes a probability of picking a hidden class b.

An observed pair (d, t) can be found, while the hidden class variable 'b' is eliminated. Converting the whole above method into a defined joint probability model yields following expressions.

$$P(d, t) = P(d)*P(t \mid d) \qquad 1$$

$$P(t \mid d) = \sum P(t \mid b) * P(b \mid d) \qquad 2$$

Expectation Maximization algorithm can be used in the model building with maximum likelihood formulation of the learning task [8]. In the EM algorithm, the posterior probabilities are computed in the E-step in equation 3.

$$P(b|d, t) = P(d)*P(t \mid d)/\sum P(t \mid b) * P(b \mid d) \qquad 3$$

PLSA Algorithm

- Inputs: term to document matrix (t ,d), t=1:n, d=1:c and the number 'k' of topics sought [11]
- Initialize arrays R1 and R2 randomly with numbers between [0,1] and normalize them row–wise to 1 [10]
- Iterate until convergence

For d=1 to c, For t=1 to n, For b=1 to k

$$R1(t,b)= R1(t,b)\sum^{n}_{d=1} \{T(t,d)*R2(b,d)/\{\sum^{B}_{b=1}R1(t,b)*R2(b,d)\}\} \quad 4$$

$$R2(b,d)=R2(b,d)\sum^{m}_{t=1}\{T(t,d)*R1(t,a)/\{\sum^{B}_{b=1}R1(t,b)*R2(b,d)\}\} \quad 5$$

$$R2(t,b)= R1(t,a)/\sum^{m}_{t=1} R1(t,b) \qquad 6$$

$$R2(b,d)= R2(b,d)/\sum^{B}_{b=1} R2(b,d) \qquad 7$$

Output: arrays R1 and R2 which hold the estimated parameters P (t | b) and P (b | d) respectively [9]

Equation 4 and 5 illustrate expectation steps in which posterior probabilities are computed from currently expected values. For initial step these estimated parameters are assigned values using a uniform random number generator which generates number 0 and

1. Equation 6 and 7 are maximization steps where initial parameters are changed from the values resulting from E step.

PLSA describes suitable probability distributions to the documents and has its basis in statistics. PLSA is interpretable with its generative model using latent classes. PLSA is given equal or better result compared to VSM in the context of information retrieval. It was also illustrated that the accuracy of PLSA can increase when the number of hidden variable increase. The problem with It’s model is that the model used to calculate the model, i.e. Expectation Maximization, can converge to a local maximum. Thus, we are not guaranteed a global optimum.

3. A WALK- THROUGH EXAMPLE

In the previous section we have briefly illustrated the algorithms that we intend to use to measure the effect of pronoun resolution on the similarity score. In this section we describe a small experiment that clearly demonstrates the effect of PR. The actual experiments are illustrated in the next section. In order to build our example we have created two set of documents. Set1 consists of three documents with no pronoun resolution while Set 2 consists of the same documents but after resolving the pronouns i.e. where the pronouns have been replaced by their corresponding nouns. Both sets are illustrated below. Human experts were asked to provide a similarity score between the documents. The same similarity score was also calculated using VSM and PLSA. The results are illustrated in Table 1.

Set 1.

D1: Ram was a gentleman. He was husband of Sita. He was the king of Ayodhya. He had two children. He had four brothers. He had three mothers. Ayodhya was the birth place of Ram. Janak was the king of Janakpuri. Ram won Lanka after killing Rawan.

D2: Ram was a gentleman. He had two children. He had four brothers. He was king of Ayodhya. Ayodhya was the birth place of Ram. Janak was the king of Janakpuri. Ram won Lanka after killing Rawan.

D3: Ram was a gentleman. He had two children. He had four brothers. Ayodhya was the birth place of Ram. Janak was the king of Janakpuri. Ram won Lanka after killing Rawan.

For both the sets we calculate the similarity score of D1 with each document in the set. The results are illustrated in Table 1. The effect of PR is quite noticeable; it changed the score by ~20% in above document set. Moreover, the scores obtained after PR are very close to the benchmark similarity scores

Table 1. Similarity scores of all documents with document D1, before and after pronoun resolution

The reason for the enhanced scores is can be found by considering the document D1 in Set1. If we observe the document we find that the term, ‘Ram’, occurs three times while the term ‘he’ (which actually refers to ‘Ram’) occurs five times. Thus, the term frequency of ‘Ram’ is three while it should have been eight. This shortcoming of the traditional document similarity measures is removed in Set2 by performing pronoun resolution, thus enhancing the similarity score. The fact that both VSM and PLSA give similar levels of enhancement indicates that the effect is not merely an artifact of the specific algorithm used for measuring the similarity score but, instead, it is due to pronoun resolution. At this stage we must admit that the example presented in this section was specifically built to demonstrate the effect of pronoun resolution. However, real world documents may show a different behavior. Therefore, in the next section we present the results on real world documents.

4. RESULT AND ANALYSIS

In order to analyze the effect of Pronoun Resolution on similarity measurement we built a corpus with different types of documents like short stories, sports news, political news, news on terrorism and scientific articles. These documents were collected from different electronic resources like websites, e-news paper etc. The number of documents of each type is: short stories - 40 sports - 80 terrorism 150 and scientific - 80.

Human were asked to provide similarity scores between pairs of documents in the collection. These scores were treated as the “ground truth” and provided a benchmark for the automated processes. AS in the previous section, the original documents were put in Set1. Set2 consisted of the same documents but after resolving all the pronouns to their respective nouns. We then calculated the similarity score of each document with all the others, separately for Set1 and Set2. Both VSM and PLSA were used for generating the scores. We thus have four automatically generated scores for each document pair. These are (VSM without PR), (PLSA without PR), (VSM with PR) and (PLSA with PR). In addition to these automatically generated scores, we also had the human generated scores for each document pair.

Since the total number of scores was very large we decided to compute the Spearman correlation coefficient between the human generated scores separately with each of the automated scores. These coefficients were computed as follows. For a given document, we consider its similarity score with all other documents, using a particular scoring technique. These numbers can be treated as an array. Thus, for every document we get an array for a given technique. The correlation coefficient is calculated between the arrays of a particular document obtained using one of the automated techniques and that obtained from the human generated scores. The overall coefficient of a particular technique is obtained by taking the average of all the coefficients obtained using that technique. These numbers then give us a basis for comparison of the distinct techniques. The values obtained using the method described above is illustrated in Table 2A and Table 2B shown below.

Table 2A: Spearman Correlation between human and system generated similarity score with before and after pronoun resolution using PLSA and VSM in average case.

Doc ID	Given Doc Id	Bench mark Similarity	Before PR VSM	Before PR PLSA	After PR VSM	After PR PLSA
D1	D1	1	1	1	1	1
D2	D1	.96	.74	.75	.95	.96
D3	D1	.94	.72	.72	.94	.94

Types of Doc	VSM BPR SP Co.	VSM APR SP Co.	PLSA BPR SP Co.	PLSA APR SP Co.
Sports	.78	.80	.79	.81
Terrorism	.68	.72	.70	.74
Scientific	.91	.91	.92	.92
Stories	.88	.93	.89	.94

Table 2B: Spearman Correlation between human and system generated similarity score with before and after pronoun resolution using PLSA and VSM in some best cases.

Types of Doc	VSM BPR SP Co.	VSM APR SP Co.	PLSA BPR SP Co.	PLSA APR SP Co.
Sports	.86	.87	.86	.87
Terrorism	.73	.77	.74	.79
Scientific	.96	.96	.96	.96
Stories	.91	1	.91	1

The effect of pronoun resolution can be observed by comparing the values in column 1 with those of column 2 and similarly by comparing the values in column 3 with those in columns 4. We can that in all the cases the correlation between the automatic technique and the ground truth improves when we introduce pronoun resolution. The second point to be noted is that the improvement in the correlation is almost independent of the actual technique used to measure the similarity i.e. for both VSM and PLSA we get similar improvement. However, the improvement does depend on the genre of the document. As can be seen, the improvement in the case of short stories is more than that observed in other types of document. The scientific articles are least affected by pronoun resolution. The reason behind this phenomenon is that short stories and news articles use pronouns more often than scientific articles. Thus, the resolution of the pronouns leads to more significant differences in the term frequency calculations and hence to bigger changes in the correlation coefficient.

It is pertinent to emphasize at this stage that we should not look upon the entries Table 2A and Table 2B as a comparison between the VSM and the PLSA techniques. What has been illustrated in the Tables are the correlations between the similarity scores obtained using the automated techniques and the human generated values. For a proper comparison between PLSA and VSM we should not look at the correlation coefficients, but rather we should look at the differences between the actual similarities scores obtained using either technique with those given by humans. The actual differences in the similarity scores obtained before and after pronoun resolution is found to be again dependent on the genre of the document. AS expected, the differences are larger for short stories and news articles and they are smaller for scientific articles.

5. FUTURE WORK AND CONCLUSION

In this paper we have measured the effect of pronoun resolution on the document similarity score. Our primary conclusion is that there is a positive effect of performing the pronoun resolution. Moreover, the effect is independent of the actual method used for measuring the similarity. Also, the effect is more pronounced for those documents that have a larger number of pronouns like short stories and news articles. The effect is less for scientific articles. The main reason of this effect is that pronoun resolution affects the term frequency count of the term document matrix. The revised counts (i.e. those obtained after pronoun resolution) lead to the improved estimates of document similarity. While the experiments were performed using English documents only, we expect the effect of pronoun resolution to be independent of the language. Moreover, since the pronoun resolution step is an offline process, so it will not add to the time complexity of the actual retrieval algorithm.

Since the present work clearly brings out the importance of performing pronoun resolution, it will be interesting to see whether other NLP processes can further improve the similarity scores. For example, we may perform a Named Entity Resolution as a preprocessing step. Each unique named entity can then be treated as a distinct term in the term document matrix. Similarly, cue phrases can e used to modify the weights of terms in a given document. Experiments along this scenario are in progress and will be presented soon. Similarly, while we have used the VSM and PLSA in the present work, it will be motivating to see whether other statistical techniques like Latent Dirichlet Allocation [13] provides further improvements.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the infrastructural support from IIIT Allahabad. One of the authors, Atul Kumar, also acknowledges the financial support from IIIT Allahabad.

7. REFERENCES

[1] Lee, D L; Huei Chuang; Seamons, K (1997) Document ranking and the Vector Space model, Software IEEE Volume 14, Issue 2 Pages 67-75, Mar/Apr (1997).

[2] Baeza –Yates, R and Riberio-Neto, B (1999) Modern Information Retrieval", Addison Wesley Longman.

[3] Salton, G; Wong, A and Yang, C S (1975) A Vector Space Model for Automatic Indexing, Communications of the ACM, vol. 18, nr. 11, pages 613 – 620.

[4] Salton, G and Lesk, M (1971)Computer evaluation of indexing and text processing", Prentice Hall, Ing. Englewood Cliffs, New Jersey. 143–180.

[5] Deerweater, S; Dumais S T; Furnas, G W; Landuar, T K and Harshman, R A (1990) Indexing by Latent Semantic Analysis, Journal of the American Society for Information science,41(6).391-407.

[6] Landauer, T K; Foltz P W and Laham D (1998)An Introduction to latent semantic analysis, Discourse Processes, vol. 25, pp. 259-284.

[7] Thomas Hofmann (1999) Probabilistic Latent Semantic Indexing, Annual ACM Conference on Research and Development in Information Retrieval, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California, United States, pp 50 – 57

[8] Thomas Hofmann (1999) Probabilistic Latent Semantic Analysis", Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.

[9] Tuomo Kakkonen, Niko Myller, Jari Timonen and Erkki Sutinen (2005)Automatic Essay Grading with Probabilistic Latent semantic Analysis, Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, pages 29-36, Ann Arbor, June (2005)

[10] Dempster P; Larid N M and Rubin D B (1977) Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, 39 1-38.

[11] University of Birmingham, School of computer science http://www.cs.bham.ac.uk/%7Eaxk/ML_PLSA.ppt

[12] Pincombe, B M (2004)Comparison of human and latent semantic analysis (LSA) judgments of pairwise document similarities for a news corpus", Defence Science and Technology Organisation Research Report DSTO–RR–0278

[13] Girolami and Kaban A ,(2003)On an Equivalence between PLSI and LDA", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 433-434, Toronto, Canada ACM Press.

[14] Turney P (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning.

[15] Leacock C and Chodorow(1998) Combining local context and Word Net sense similarity for word sense identification,In WordNet an Electronic Lexical Database. The MIT Press.

[16] Wu Z and Palmer M (1994)Verb semantics and lexical selection, Proceedings of the Annual Meeting of the Association for Computational Linguistics.

[17] Rocchio J(1971)"Relevance feedback in information retrieval, Prentice Hall, Ing. Englewood Cliffs, New Jersey.

[18] Mihalcea R, Corley C and Strapparava C(2006) Corpus-based and Knowledge-based Measures of Text Semantic Similarity, AAAI'06, pp 775-780.

[19] Hammouda K M, Kamel M S (2004)Document similarity using a Phrase Indexing Graph Model, Knowledge and Information Systems Springer –Verlag London 6:710-727(2004)

[20] Xu R, Wunsch II D (2005) Survey of clustering algorithm. IEEE Trans Neural Netw 16(3):645-678.

[21] Vivekanandan K and Suguna J(2008)Inferring Document Similarity using the Fuzzy measure, Medwell Journals - Asian Journal of Information Technology 7 (1):1-5.

[22] Wan X and Peng Y(2005)The earth mover's distance as a semantic measure for document similarity, Proceedings of the 14th ACM international Conference on Information and Knowledge Management Bremen, Germany, October 31 - November 05, CIKM '05. ACM Press, New York .