A Term Association Inference Model for Single Documents: A Stepping Stone for Investigation through Information Extraction

Sukanya Manna and Tom Gedeon

Department of Computer Science,
The Australian National University, Canberra, ACT, Australia
{sukanya.manna,tom.gedeon}@anu.edu.au

Abstract. In this paper, we propose a term association model which extracts significant terms as well as the important regions from a single document. This model is a basis for a systematic form of subjective data analysis which captures the notion of relatedness of different discourse structures considered in the document, without having a predefined knowledge-base. This is a paving stone for investigation or security purposes, where possible patterns need to be figured out from a witness statement or a few witness statements. This is unlikely to be possible in predictive data mining where the system can not work efficiently in the absence of existing patterns or large amount of data. This model overcomes the basic drawback of existing language models for choosing significant terms in single documents. We used a text summarization method to validate a part of this work and compare our term significance with a modified version of Salton's [1].

Keywords: Information retrieval, investigation, Gain of Words, Gain of Sentences, term significance, summarization.

1 Introduction

Information retrieval (IR) deals with text analysis, text storage, and the retrieval of stored records having similarity between them [2]. Among various IR models, vector based model is the significant one assigning weights based on the discriminative powers [3]. Inverse Document Frequency is the most common language model. But there are also modifications of the above concept into inverse sentence frequency and inverse term frequency, which all work over a large corpus to find a solution to the problem where document space language models do not work [4]. There are situations when the user query is not the only desired need but the relations between different contexts within a single text, which provide an insight into the semantic relations, might be of interest in some specific applications like official investigations, or counter terrorism, text summarization [5], question answering systems [6] and so on.

There are different computational models for natural language discourse structures, which are mainly used for summarization and question answering systems [7],[8], [9], [10]. In [11], the authors generate intra-document semantic hyperlinks and characterize the structure of a text based on the intra document linkage pattern. Again the concept of

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 14–20, 2008.

Latent Semantic Analysis [12] exploits knowledge induction and representation. A related concept to our work was analyzed by Rocha [13], where he presented keyword semantic proximity and its semi-metric behaviour in a recommendation system TalkMine to advance adaptive web and digital library technology.

IR following conventional predictive data mining techniques has proved to be ineffective in handling cases where there are no previous patterns of data available [14]. If we consider the act of terrorism, we do not find any similar indicia. With a relatively small number of attempts every year and only one or two major terrorist incidents every few years- each one distinct in terms of planning and execution- there are no meaningful patterns that show what behaviour indicates planning or preparation for terrorism. So, it is preferable to handle these types of scenarios with subjective data analysis or computational linguistic technologies to exploit the semantic and syntactic structure of texts.

Almost every document has some hierarchical structure concerning the importance of the words or concepts occurring in it [15]. The basic idea of linking the terms (entities + significant keywords) in a document is based on their frequency of occurring together in different paragraphs or sentences, presuming them to have some relationship. This approach does not require any previous knowledge about the data pattern. It is based on the degree of linkages found between different terms and brings out the relevant ones.

2 Motivation

In the previous section we have already mentioned that predictive data mining is not that useful to analyze cases like terrorism [14], or social crimes. Trained officials need to analyze every witness statements to find some clues to assume a possible solution to solve a legal problem. The basic objective of our work is to enhance the performance of these people and make their work easier in getting a solution.

There are several works related to information extraction, but the established models [3], [12], and [13] mainly deal with huge corpora for their analysis. Hence, it is challenging to work with a single document or very few documents to extract the most important facts and create a possible network to find patterns between different discourse segments within the text.

3 Term Significance Models

3.1 Modification of Salton's Indexing Method for Choosing Significant Terms in Single Documents

In this section we have modified Salton's [1] term discrimination model in such a way so that the documents in his model refers to the sentences in our version. Instead of calculating the similarities between the document pairs, we calculated here the similarity between the sentence pairs respectively. Our main aim of calculating the discrimination value was to identify the significant terms.

Let sentences be the discourse structure in this case. So, similarity between sentences is calculated by,

$$sim(s_i, s_j) = \sum_{k=1}^{t} w_{ik} w_{jk} \tag{1}$$

where, $t = no.\ of\ attributes\ (\ or\ terms)$, w refers to the binary weights, i.e.,

$$w = \begin{cases} 1, & if\ \ the\ term\ is\ \ present\ in\ the\ sentence \\ 0, & otherwise \end{cases} \qquad (2)$$

and $i, j \in sentences$.

The average similarity between the sentence pairs is calculated by,

$$sim_{avg} = K \sum_{\substack{i=1, j=1 \\ i \neq j}}^{n} sim(s_i, s_j) \qquad (3)$$

Now, consider the original sentence collection with the term k removed from all the sentence descriptions and let sim_{avg}^{k} be the average sentence pair similarity in that case.

So, *discrimination value (DV)* can be computed as,

$$DV = (sim_{avg}^{k} - sim_{avg}) \qquad (4)$$

According to Salton [1], if $DV > 0$, it refers to good discriminators and if $DV < 0$, it refers to bad discriminators.

3.2 Our Approach: Gain of Words (GOW)

We present here a method, whose major purpose is to discriminate between the significant and non significant terms (or words). As a preprocessing step, we have initially considered all words from the document including the stop words.

Now, let n be the no. of words/ terms considered. Let S be the vector of sentences present in the document. So, we calculated the gain of words by,

$$GOW = \frac{\sum f_{ij}}{\sum i} \times \sum w_{ij} \ , \qquad (5)$$

Where, f_{ij} is the frequency of the term (no. of occurrences) j in the sentence i and w is the weight as mentioned in the previous model.

Words having very high *GOW* values are discarded, maintaining a threshold of *0<GOW<10.*

4 Sentence Extraction: *Gain of Sentences (GOS)*

Gain of Sentences, refer to the value which signifies the importance of sentences in a document. The greater the value, higher is the importance. Before computing this, in the preprocessing stage, we discarded all the stop words. As mentioned above, let n be the no. of words / terms considered. Let S be the vector of sentences present in the document.

So, we compute the gain of sentences by,

$$GOS = \frac{\sum f_{ij}}{\sum j} \times \sum w_{ji} \ , \qquad (6)$$

Where, f_{ij} is the frequency of the term (which means no. of occurrence of the term) j in the sentence i and w is the weight as mentioned in the previous model.

This concept is also used for summarizing a document as it ranks the sentences as per their importance.

5 Experimental Results

In this section we illustrate the experimental results related to the methods discussed in the previous section.

We used the CST data set [16], related to a Milan plane crash. There are multiple single texts in the data set. Since, we focus on single documents, we used each of those for analysis.

Gain of Words: We explained the significance of using G*ain of Words* (GOW) in the previous section. Using GOW, we can eliminate the unwanted words, at the same time keep the possible important words including the entities (the ones generally obtained using named entity extractors). Here we have taken ten words, randomly chosen from the files separately. The tables below show the nature of results obtained using Salton's term significance measure on a single document as well ours. It clearly shows that, for certain words, it gives some meaningful results showing that negative discriminative value, signifying that those words are poor terms. But on the other hand, it cannot differentiate between the good words also. The zeros in the tables show that it cannot identify the terms. Our result overcomes this drawback. The value of the gain computed easily helps us to identify the words between useless, useful and less useful. When the gain values are very large, it shows that the words are useless.

Table 1 and Table 2 illustrate the term significance based on two different methods. It is clearly seen in fig.1 that the highest value for the DV is 0. It is just capable of discarding the most useless terms. The words like "the", "in" (shown in table 1) are the stop words which can be discarded using both the methods. But words like "crash", "plane", "Milan" bear meaningful content, but can be identified by GOW method, not with DV.

Table 1. Comparison between two term significance methods

Document 1		
Words	DVof Salton's method	GOW
the	-0.462	13.847
in	-0.462	13.154
plane	-0.038	0.692
building	-0.077	1.231
crash	-0.038	0.692
april	0	0.077
skyscraper	-0.013	0.308
milan	-0.013	0.308
cnn	-0.0123	0.462
bombing	0	0.077

Table 2. Comparison between two term significance methods

Document 2		
Words	DV of Salton's method	GOW
Are	-0.035	0.842
smoke	-0.006	0.211
police	-0.006	0.211
people	-0.006	0.211
milan	-0.006	0.211
scene	-0.018	0.474
pirelli	-0.018	0.474
italian	-0.018	0.474
from	-0.018	0.474
work	-0.018	0.474

The DV identifies all words as useless, except for “april” and “bombing” and identifies these as almost useless. Our technique also identifies these two as almost useless, but also identifies stop words and useful words and clearly differentiates them.

We basically maintained a threshold of *0<GOW<10* approximately to choose the words. But there is some noise in out data also.

In table 2, the words like “are”,” from” fall within the threshold limit we have chosen. So these words could not be identified

Gain of Sentences: The Gain of sentences (GOS) is another useful method we present here. The basic target of this part is to analyze and extract the important regions of the text so it partly behaves as a summarization. We had to do some preprocessing before obtaining GOS. Using the threshold mentioned above, we selected the possible significant words from the text. But in order to reduce the noise, we removed the stop words from this new set of words. After this we ran our simulation to obtain GOS. We used the MEAD [17] summarization tool to compare our method. We present here the nature of summaries extracted using this tool as well as with our method. Since GOS creates sentence importance in the document, we here present the five most important sentences to see how far it holds with the MEAD’s process.

For Document 1:

MEAD summarization:

CNN.com - Plane hits skyscraper in Milan - April 18 2002 CNNenEspanol.com A small plane has hit a skyscraper in central Milan setting the top floors of the 30-story building on fire an Italian journalist told CNN.

The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. 1450 GMT on Thursday said journalist Desideria Cavina.

U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. 2002 Cable News Network LP LLLP.

Our Approach: GOS

CNNenEspanol.com A small plane has hit a skyscraper in central Milan, setting the top floors of the 30-story building on fire, an Italian journalist told CNN.

U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. 2002 Cable News Network LP, LLLP.

The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. (1450 GMT) on Thursday, said journalist Desideria Cavina.

Italian TV says the crash put a hole in the 25th floor of the Pirelli building, and that smoke is pouring from the opening.

Many people were on the streets as they left work for the evening at the time of the crash.

For Document 2:

MEAD summarization:

CNN.com - Plane hits skyscraper in Milan - April 18 2002 CNNenEspanol.com A small plane has hit a skyscraper in central Milan setting the top floors of the 30-story building on fire an Italian journalist told CNN.

The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. 1450 GMT on Thursday said journalist Desideria Cavina.

I heard a strange bang so I went to the window and outside I saw the windows of the Pirelli building blown out and then I saw smoke coming from them said Gianluca Liberto an engineer who was working in the area told Reuters.

U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. 2002 Cable News Network LP LLLP.

Our Approach: GOS

"I heard a strange bang so I went to the window and outside I saw the windows of the Pirelli building blown out and then I saw smoke coming from them," said Gianluca Liberto, an engineer who was working in the area told Reuters.

TV pictures from the scene evoked horrific memories of the September 11 attacks on the World Trade Center in New York and the collapse of the building's twin towers.

CNNenEspanol.com A small plane has hit a skyscraper in central Milan, setting the top floors of the 30-story building on fire, an Italian journalist told CNN.

U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. 2002 Cable News Network LP, LLLP.

The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. (1450 GMT) on Thursday, said journalist Desideria Cavina.

The alignment of the sentences we presented might vary from the MEAD summarization. Many of the documents include the same sentences as news sources import the same sentences into different documents. We have presented here the sentences based on their importance in the document. Clearly, our summaries are qualitatively equivalent to the MEAD summarizations.

6 Conclusion

This work is a two way approach of term association where we find the significant words as well as extract the important sentences from a text. It is a simple method based on the syntactic appearances of the terms/ words in a single document. It is very useful to analyze the cases where no predefined data pattern is available. We have also shown that a classic method which has been used successfully for term extraction fails to work when there is a single document or very few documents. Though we have seen that the performance of this model is better, but still we need to improve this in order to get rid of the noise.

References

1. Salton, G.: A Theory of Indexing. In: Regional Conf. Series in Applied Mathematics, Philadelphia, Pennsylvania (1975)
2. Salton, G., Fox, E.A., Wu, H.: Extended Boolean Information Retrieval. Comm. of ACM 26(12), 1022–1036 (1983)
3. Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. Inf. Processing and Management 24(5), 513–523 (1988)
4. Blake, C.: A Comparison of Document, Sentences, and Term Event Spaces. In: Proc. of 21st intl. Conf. on Comp. Linguists and 44th Annual Meeting of the ACL, pp. 601–608 (2006)
5. Zhang, Z., Goldensohn, S.B., Radev, D.R.: Towards CST-Enhanced summarization. In: Eighteenth national conf. on Artificial intelligence, pp. 439–445 (2002)
6. Katz, B., et al.: START, Natural Language question answering system (1993)
7. Zhang, Z., Otterbacher, J., Radev, D.: Learning Cross-document structural Relationships using Boosting. In: CIKM (2003)
8. Grosz, B.J., Sidner, C.L.: Attention, Intentions, and the Structure of Discourse. Comp. Linguistics 12(3), 175–204 (1986)

9. Radev, D.R.: A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure. In: Proc. of the First SIGdial Workshop on Discourse and Dialogue (2000)
10. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: towards a functional theory of text organization. Text 8(3), 243–281 (1988)
11. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Inf. Processing & Management 33, 193–207 (1997)
12. Landauer, T.K., Foltz, P.W., Laham, D.: An introducition to Latent Semantic Analysis. Discourse Process 25, 259–284 (1998)
13. Rocha, L.M.: TalkMine. A Soft Computing Approach to Adaptive Knowledge Recommendation. In: Loia, V., Sessa, S. (eds.) Soft Computing Agents: New Trends for Designing Autonomous Systems. Series on Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer (2001)
14. Jonas, J., Harper, J.: Effective Counterterrorism and the Limited Role of Predictive Data Mining. Policy Analysis 584, 1–12 (2006)
15. Gedeon, T.D., Koczy, L.T.: Hierarchical co-occurrence Relations. In: Proc. Systems, Man, and Cybernetics, vol. 3, pp. 2750–2755 (1998)
16. Radev, D., Otterbacher, J., Zhang, Z.: CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In: Proc. of LREC 2004 (2004)
17. Radev, D., et al.: MEAD - a platform for multidocument multilingual text summarization. In: LREC, Lisbon, Portugal (2004)