LINGUISTICALLY ENHANCED CLUSTERING OF TECHNICAL PUBLICATIONS

Mahmoud Gindiyeh[1], Gintarė Grigonytė[1,2], Johann Haller[1] and Algirdas Avižienis[2]
[1] *Saarland University, Saarbrucken, Germany*
[2] *Vytautas Magnus University, Kaunas, Lithuania*

Keywords: Linguistic analysis, Information retrieval, Clustering.

Abstract: Organizing documents and performing search is a common but not a trivial task in information systems. With the increasing number of documents, it is becoming crucial to automate these processes. Clustering is a solution for organizing large amount of documents. In this article we propose a method of improving document retrieval that was implemented in RKB Knowledge Base. Our method heavily relies on linguistic analysis, which aims to identify document specific noun phrases. We apply an adjusted hierarchical clustering algorithm for learning clusters of documents.

1 INTRODUCTION

Our work is motivated by the need to improve the search feature of scientific publications in Resilience Knowledge Base (RKB) - online knowledge base (Glaser et.al., 2008). Instead of using a simple pattern based technique, which looks for given word occurrences in text or its meta-data structures, it is possible to find documents topically similar to a given document.

For that we need to train a classifier which will be used to identify a group of the documents that are highly related to a given one. The RKB knowledge base online interface presents over 50 million of interlinked items: research projects, researchers, and publications. Users are able to traverse the datascape by altering selection topics and choosing search results. For instance, when searching for similar publications to ones already known, the user locates a known title and is presented with a list of highly related publications.

When performing this kind of search, the aim is to return only a highly relevant result. The user expects to find similar publications within the first top 5-10 list. One possible solution how to optimize search results is document clustering (Kouomou et.al., 2005), (Gelbukh et.al., 1999).

Instead of a simple 'bag of words' method, some approaches of document clustering relies on citation analysis, such as (Huang et.al., 2004), or (Joerg, 2008).

However, as suggested by similar to ours researches, the vector space model for text retrieval is giving better results if the indexing space is based on linguistic features such as WordNet synsets (Gonzalo et.al., 1998) or noun phrases (Hatzivassiloglou, 2000) instead of a plain statistics of word forms. Other similar research include (Tikk et.al., 2007) and (Zheng et.at, 2009).

2 THE METHOD

The clustering method that was used in this experiment is based on combining Pearson's correlation values as similarity distance measures and applying a hierarchical clustering algorithm. In order to acquire distance measures we use numeric values that show topical importance of noun phrases (NPs) in a particular document. To calculate these numeric values we perform morphological and syntactical analysis of documents and use a technical thesaurus.

Stepwise our method can be divided into 4 phases:

1. Identification of the topically specific NPs in the documents.
2. Creation of feature representations (NPs and their weights) for each document.
3. Calculation of a similarity degree and population of the similarity matrix.
4. Applying the clustering algorithm on the basis of the similarity matrix.

Gindiyeh M., Grigonyte G., Haller J. and Avižienis A. (2009).
LINGUISTICALLY ENHANCED CLUSTERING OF TECHNICAL PUBLICATIONS.
In *Proceedings of the International Conference on Knowledge Management and Information Sharing*, pages 324-327
DOI: 10.5220/0002308703240327

Firstly, each document is linguistically analyzed. Tasks include a lemmatisation, a part of speech tagging, and a partial semantic tagging. We have used MPRO software (Maas et.al., 2009) for that matter. Consequently, the noun phrases are marked in each document.

As a following step, the importance of each noun phrase – a weight – is calculated. The NPs are weighted by means of the thesaurus (in our case, we have used the English version of FIZ thesaurus (FIZ Technik, 2000)) and the results of linguistic analysis. The weight is calculated according to:

- the frequency of NP in the particular document as well as in other documents;
- the status of the NP in relation to the thesaurus: whether it is a hypernym, synonym or hyponym, or has no correspondence to the thesaurus;
- the number of semantic classes allocated to the particular NP during the linguistic analysis;
- the number of semantic classes allocated to the document;
- and the position of the NP in the document (beginning, end, etc.).

A detailed description of the formula we have used, is implemented in AUTINDEX software, described in (Haller and Schmidt, 2006). This approach allows us to pick only document topical NPs, as we take the context into consideration.

NPs and their weights are used for building feature vectors for each document. Subsequently, each document is represented as a vector in vector space R^N whose elements are the NPs and their weights. For example, a document vector appears as following:

```
D= (computer system [100]; research
and development [87]; error [28];
encryption protocol [27]; project
planning [21]; security [14])
```

We assume the vector space $V = (V_1, V_2, \ldots, V_j, \ldots, V_N)$, where V_j is the j-th document characterizing feature vector.

The matrix of documents has columns which are feature vectors (documents) and rows – NP_i which refers to NP's weight representing each document.

Finally, the similarities between feature vectors $V' = (V'_1, V'_2, \ldots, V'_j, \ldots, V'_n)$ are calculated. We have chosen to express the similarity through the Pearson's correlation coefficient. Correlation indicates the strength and direction of a linear relationship between two variables. The coefficient is represented in the interval [-1,1]. Therefore it is simple to decide whether given variables are similar or not, i.e. from non-related (-1) to matching (+1).

The following simplified correlation rule was used:

$$Corr_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{(Var(Y))}} \quad (1)$$

In order to enhance the contrast of similarity values we have calculated similarity values for the second time – again applying Pearson's correlation rule but this time not on term weights as before, but on similarity values from the previous step. The similarity values obtained in this way are distributed differently. The contrast between most similar documents and not-so-similar documents is a lot higher, as shown in Table 1.

When performing document clustering the aim is to divide a quantity of documents into topic-specific groups. These groups are not known in advance.

The hierarchical clustering algorithm we have applied is similar to the one described in (Johnson, 1967), or (Manning and Schütze, 1999). In addition, during the experiment we have added constraints of disjoint and joint clusters, and have extended similarity matrix by calculating the correlation of correlation between documents.

3 THE EXPERIMENT

Our experiment set was around 2500 scientific articles from the domain of computer dependability and security.

We have performed experiments according to following settings:

- Disjoint clusters vs. joint clusters
- Similarity can be calculated either once or twice.

In our clustering algorithm, we have used the threshold value, which was selected by experts of the domain. The motivation of choosing threshold was that a smaller threshold value delivers too big clusters, i.e. an irrelevant document is more likely to be assigned to a cluster.

On the other hand, when the threshold value is set too high, clusters tend to be very small which is undesirable for the purpose of searching for publications in RKB. As a side effect, quite many documents remain unclustered. Results of the experiment are presented in Figure 1, and Table 2 – columns represent different experimental settings, i.e. 1-pass correlation and joint clusters, 1-pass correlation and disjoint clusters, 2-pass correlation and joint clusters, and 2-pass correlation and disjoint clusters.

Table 1: Similarity values for the article "Feedback Bridging Faults": similarities in % - the list of the 10 most similar documents, calculated by 1-pass correlation and 2-passs correlation method.

1-pass correlation similarity values		2-passs correlation similarity values	
FEEDBACK BRIDGING FAULTS	%	FEEDBACK BRIDGING FAULTS	%
1) bridging and stuck-at faults	66	1) design of fault-tolerant clocks with realistic failure assumptions	98
2) on undetectability of bridging faults	61	2) efficient distributed diagnosis in the presence of random faults	98
3) test generation for mos complex gate networks	56	3) software schemes of reconfiguration and recovery	98
4) a nine-valued circuit model to generate tests for sequential circuits	52	4) towards totally self-checking delay-insensitive systems	97
5) sharpe 2002: symbolic hierarchical automated reliability and performance	51	5) on partial protection in groomed optical wdm mesh networks	97
6) concurrent fault diagnosis in multiple processor systems	51	6) test generation for mos complex gate networks	95
7) the algebraic approach to faulty logic	51	7) concurrent fault diagnosis in multiple processor systems	86
8) a two-level approach to modeling system diagnosability	51	8) bridging and stuck-at faults	86
9) design of fault-tolerant clocks with realistic failure assumptions	51	9) computer-aided design of dependable mission critical systems	66
10) a model of stateful firewalls and its properties	51	10) efficient byzantine-tolerant erasure-coded storage	56

Table 2: The number of clusters learned from 2469 documents.

	1-pass Corr joint	1-pass Corr disjoint	2-pass Corr joint	2-pass Corr disjoint
#Clusters	287	199	149	88
#Unclustered documents	399	538	33	52

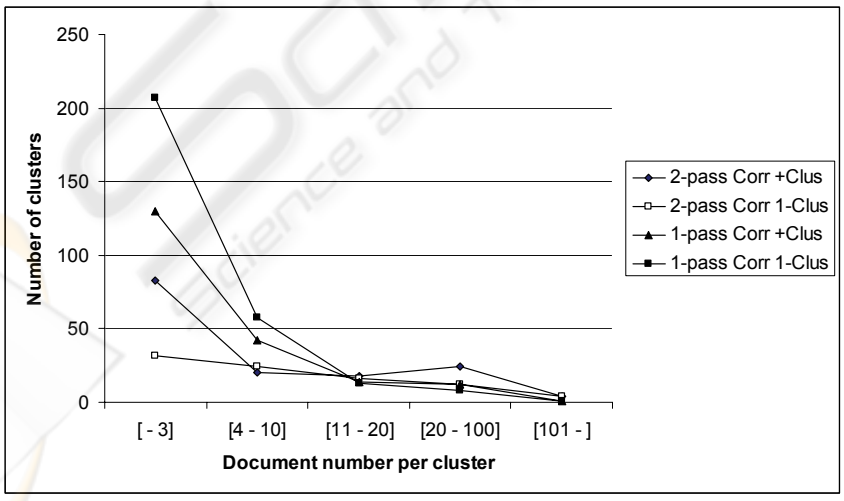

Figure 1: 4 experiments - distributions of document number per cluster.

For the purposes of RKB Knowledge Base, the method that is able to assign the majority of the documents into clusters is considered better, as all the publications in the knowledge base belong to the same domain. A distribution with a lot of small clusters, i.e. with 2 documents, or large clusters, i.e. 50, 100 and more, is unwanted, as it is not optimal for search purposes. Considering these constraints, the 2-pass correlation and joint clusters method was the most appropriate.

4 CONCLUSIONS

The evaluation of 10% of the set of documents used for the experiment was performed manually by experts of the domain. The most relevant clusters were created with 2-pass correlation and joint clusters method. The extrinsic measures of clustering were 0,87 of purity and 0,22 of entropy. The evaluation on Reuters data set showed 0,62 of purity and 1,44 of entropy (see Table 3.). These differences appear because of the genre and topic of texts present in Reuters data set – general language corpus. One important aspect of our methodology is using technical thesaurus to assign the importance weight to NPs found in text.

Table 3: Results of clustering methodology applied on technical documents and general language texts.

Text genre	Purity	Entropy
Technical documents	**0,87**	**0,22**
General texts	0,62	1,44

The results of this experiment were applied in RKB Knowledge Base. When viewing a particular publication, RKB Knowledge Base provides a list of most relevant publications.

ACKNOWLEDGEMENTS

The authors wish to thank Hugh Glaser and Ian Millard of Southampton University for their advice and cooperation regarding the Resilience Knowledge Base. This research has been supported in part by EC IST contract no. 026764, Network of Excellence ReSIST (Resilience for Survivability in IST). Gintare Grigonyte was supported by DAAD (Deutscher Akademischer Austauschdienst) grant A/07/92317.

REFERENCES

Gelbukh, A.F., Sidorov, G., Guzmán-Arenas, A. 1999. Use of a Weighted Topic Hierarchy for Document Classification. In *Proceedings of the 2nd international Workshop on Text, Speech and Dialogue* V. Matousek, P. Mautner, J. Ocelíková, and P. Sojka, Eds. Lecture Notes In Computer Science, vol. 1692. Springer-Verlag, London, 133-138.

Glaser, H., Millard, I., Jaffri, A. 2008. RKBExplorer.com: A Knowledge Driven Infrastructure for Linked Data Providers. *The Semantic Web: Research and Applications*, Springer, 797-801.

Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J., 1998. Indexing with WordNet synsets can improve Text Retrieval. *In proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP*.

Haller, J., Schmidt, P. 2006. AUTINDEX - Automatische Indexierung. *Zeitschrift für Bibliothekswesen und Bibliographie: Sonderheft 89,* Klostermann, Frankfurt am Main, 104-114.

Hatzivassiloglou, V. , Gravano, L., Maganti, A. 2000. An investigation of linguistic features and clustering algorithms for topical document clustering. *In proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 224-231.

Huang S., Xue G., Zhang B., Chen Z., Yu Y., Wei-Ying Ma 2004. TSSP: A Reinforcement Algorithm to Find Related Papers, *Proceedings of the Web Intelligence, IEEE/WIC/ACM*, p.117-123.

Johnson S. C. 1967. Hierarchical Clustering Schemes. *In Psychometrika*, 2:241-254.

Joerg B. 2008. Towards the Nature of Citations, *In poster proceedinds of FOIS 2008*, 31-36.

Kouomou, A., Berti-Équille, L., Morin, A. 2005. Optimizing progressive query-by-example over pre-clustered large image databases, *In proceedings of the 2nd international workshop on Computer vision meets databases*, Baltimore, USA.

Mass, H.D., Rösener, C., Theofilidis, A. 2009. Morphosyntactical and semantic analysis of text: The MPRO tagging procedure. *Forthcoming: SFCM 2009 workshop on Systems and Frameworks for Computational Morphology*, Zürich, Switzerland.

Manning, C., Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press Cambridge, MA.

FIZ Thesaurus Technik und Management. Hierarchisch strukturiertes Fachwortverzeichnis. 2000. *FIZ-Technik Presse-Information.* Frankfurt.

Tikk, D., Biro, G., Szidarovszky, F., Kardkovacs, Z., Lemak, G., 2007. Topic and language specific internet search engine. *In journal Acta Cybernetica*, vol. 18.2, 279-291.

Zheng, H., Kang, B., Kim, H., 2009. Exploiting noun phrases and semantic relationships for text document clustering, *In Information Sciences,* vol. 179.13, 2249-2262.