ALIYA NUGUMANOVA
East Kazakhstan State Technical University, Kazakhstan

DINARA ISSABAEVA
Kumash Nurgaliev College, Kazakhstan

YERZHAN BAIBURIN
East Kazakhstan State Technical University, Kazakhstan

AUTOMATIC GENERATION OF ASSOCIATION THESAURUS BASED ON DOMAIN-SPECIFIC TEXT COLLECTION

Abstract:

The given work examines distributive approach for automatic generation of the associative thesauri of a definite domain. Distributive approach is based on assumption that presence of associative link among terms of the domain is defined by the statistics of their co-occurence in thematically related discources. The advantage of distributive approach is defined by the fact that it uses raw basic material (for example collection of documents of the domain) and it does not use additional knowledge about the domain. Distributive approach is supported only by mathematical apparatus of statistics and does not take into account neither lexical nor semantic information, that is why this approach let cover extensive lexical space of terms. However it leads to the main shortcoming of the approach, i.e. it produces excessive amount of “unnecessary” links among words which are less informative from utilitarian point of view. For solving set problems in the given work it is suggested to use special approach represented by combination of methods of distributive statistics, latent semantic analysis and graph theory.

Keywords:

LSA, thesaurus, chi-square test, graph

JEL Classification: C80

1 Introduction

Thesaurus is a set of items (words or phrases) of some domain, connected by definite semantic relations (Jing & Croft 1994). Thesauri are divided into two types according the method of generation, i.e. thesauri created manually involving experts' work and automatically generated ones. Manual method is characterized by high labor intensiveness, so methods of automatic generation have a high relevance.

Association thesaurus is a special kind of common thesaurus, the most easily submitted to the automatic generation. There is only one kind of semantic relation between words in association thesaurus, i.e. association link. Association thesauri are widely used for solving many problems of information retrieval, i.e. query extensions, text categorization, machine translation etc. (Sinopalnikova 2004; Sérasset & Chevallet 2004; Lee et al. 2007; Ito et al. 2008; Wang et al. 2008; Pinto et al. 2008).

One of the approaches used for automatic generation of association thesaurus is co-occurrence analysis (Schutze & Pedersen 1997; Ding & Engels 2001; Morita et al. 2011). This approach is based on assumption that presence of semantic relation between two words is defined by the statistics of their co-occurrence in similar contexts (Firth 1957; Lindén & Piitulainen 2004). The advantage of co-occurrence analysis is that it uses raw basic material without additional knowledge about the domain. However, this approach has a major drawback. Data sparseness and information noise are the two factors that degrade the statistical method of co-occurrence analysis (Lindén & Piitulainen 2004; Ido et al. 1999; Zhang & Sumita 2007).

The aim of this paper is to research possible way of smoothing over the shortcomings arising from sparseness and noisiness of data. For solving set problems, we suggest to use an approach based on latent semantic analysys (LSA) (Deerwester et al. 1990; Dumais 1990). The essence of this approach consists in approximation of a matrix describing occurence of words in considered texts by a matrix of a smaller rank by means of singular value decomposition (SVD) (Deerwester et al. 1990). SVD decreases the significance of random semantic links, and increases the significance of essential ones (Dumais 1990, De Lathauwer et al. 2000). As a result, it reveals latent semantic links and decreases sparseness and noisiness in the considered texts.

According to the set goal, the structure of the paper is as follows. The 2nd part includes description of a way of the word-by-documents co-occurrence matrix construction and methods used for it. The 3rd part contains method of LSA for extracting essential semantic links between words. The 4th part gives description of method of forming of associative thesaurus, using extracted semantic links. The 5th part gives description experimental (test) section of the work. Review of the experimental results and main summary can be found in the 6th part.

2 The co-occurrence matrix construction

Let us assume that there is a domain-specific collection of documents. It is necessary to build a word-by-documents co-occurrence matrix of this collection. The co-occurrence matrix is defined as a matrix of frequencies of occurrence of the collection's words in the collection's documents. Since the purpose of this study is to build a domain-specific thesaurus, the co-occurrence matrix should be formed not from all the collection's words, but only those related to the domain. Therefore, the first step of the co-occurrence matrix construction is to extract collection's words, which related to the domain.

Many effective methods for solving this problem were developed, and the most simple and robust ones are based on the statistics of words (Manning et al 2008). The main idea of such methods is to compare a word's distribution between positive and negative samples of a category (domain). Therefore the use of these methods requires that the document collection contains both positive and negative samples of documents. Positive sample is a set of documents related to the domain, and negative one is a set of documents unrelated to the domain.

In this paper, we use a method based on Pearson's chi-squared test. This very common statistical method tests null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution (Lancaster & Seneta 1967). The test operates on the expected (theoretical) and observed (empirical) frequencies of the events. Expected frequency is calculated on the basis of theoretical distribution of the events. Observed frequency is obtained in the sample. The value of the test-statistic is:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o and f_e are the observed and expected frequencies, respectively, and the summation is over all the possible combinations of outcomes.

In our case, the null hypothesis is that there is no connection between some word and the domain. Accordingly, the alternative hypothesis is that there is some semantic link between the word and the domain, i.e. the word is related to the domain. In order to investigate the hypothesis, it is necessary to calculate the statistics of the use of each word in the collection and fill in the contingency table of dimension 2x2, as shown below.

Table 1: The contingency table

Number of documents…	… which contain this word	… which do not contain this word	Total
… related to the domain	*A*	*C*	*TSD=A+C*
… unrelated to the domain	*B*	*D*	*TSN=B+D*

Total	*A+B*	*C+D*	*A+B+C+D*

There are four options of the outcomes with the observed frequencies *A*, *B*, *C* and *D*. If a word is not related to the domain, it should be equally distributed among the domain texts, and among other texts, i.e. it should be present in the same ratio (A+B)/(A+B+C+D) as in the entire collection, and should be absent in the same ratio (C+D)/(A+B+C+D) as in the entire collection. Consequently, we can find the observed frequencies of the outcomes by multiplying corresponding numbers of texts (A+C) and (B+D) by these ratios. So, the formula to calculate the Chi-square test is:

$$\chi^2 = \frac{(A+B+C+D)(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)} \quad \text{(1)}$$

Thus, for each word of the collection the values of A, B, C, D should be found, and then Chi-square value should be calculated. If the Chi-square value is greater than the critical value, the relation between word and domain exists. A special table based on the number of degrees of freedom and a significance level (error probability) determines the critical value (Lancaster & Seneta 1967). The number of degrees of freedom is defined as $\mathrm{df} = (\mathrm{R} - 1)(\mathrm{C} - 1)$, where R and C the numbers of rows and columns of contingency table. For this task the number of degrees of freedom is 1. When a significance level is 1% and the number of degrees of freedom is 1, the critical value is 6.6.

After domain-specific words selection it is necessary to form co-occurrence matrix “words-by-documents”, containing information about distribution of the selected words in positive documents of the collection.

3 The co-occurrence matrix decomposition for extracting essential association links between words

It’s obvious that the constructed co-occurrence matrix is characterized by noisiness and sparseness. We used the method of singular value decomposition to improve the quality of the data represented by the matrix. According to the theorem of singular decomposition (De Lathauwer et al. 2000), we can present the obtained matrix in the form of product of three matrices:

$$A = U \times S \times V^T \quad \text{(2)}$$

where A is an initial matrix of m×n dimension, U and V are orthogonal matrices of m×n and n×n dimensions respectively, S is a diagonal matrix of n×n dimension, whose elements are decreasingly ordered on the main diagonal. Terms $\mathrm{s_i}$ called singular numbers of the matrix are equal to arithmetic values of square roots of corresponding eigenvalues of matrix $\mathrm{AA^T}$.

Such decomposition possesses a property that if in matrix S only k largest singular values are left, and in matrixes U and V – only columns corresponding to these values and rows are left then product of obtained matrices (new matrix A') will be the best approximation of rank k of the initial matrix A (De Lathauwer et al. 2000). The main issue of the LSA is that singular decomposition via decrease of rank of a matrix allows obtaining a new matrix expressing associative links between words and texts more evidently (Dumais 1990).

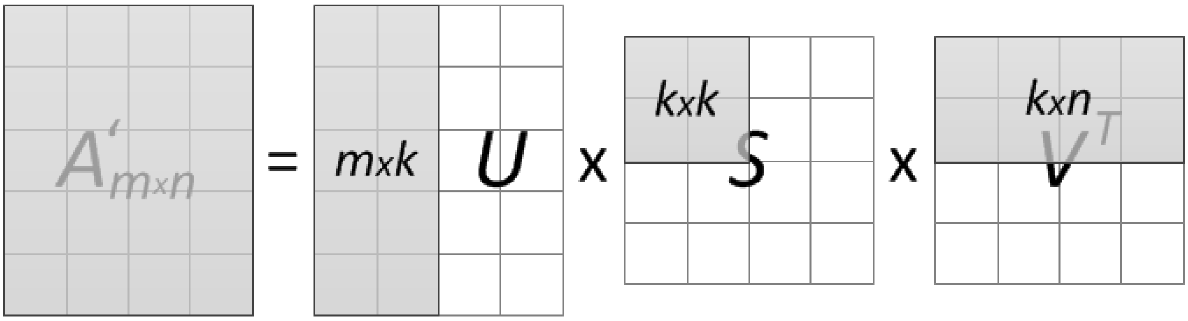

Figure 1: Approximation after SVD

4 Associative thesaurus construction using extracted semantic links

Thus, arising from the new approximated co-occurrence matrix *A'*, we can define a presence of association link between a word and a document, a word and a word, a document and a document by calculating distance between corresponding vectors. We can use a cosine measure for calculating distances:

$$c = \cos(\bar{x},\bar{y}) = \frac{\bar{x}\cdot\bar{y}}{|x|\cdot|y|} \tag{3}$$

where c is cosine measure; x, y are rows, presenting words in the new matrix. These pairwise distances express the strength of the links between words. Therefore, numbers within the range from 0 to 1 measures the links between words; the higher cosine measure is, the stronger the link is. Thus, maximum strength of link may be equal to 1, and minimum strength may be equal to 0. To extract essential links it is necessary to select only those pairs of words having the link strength more than some threshold value (for example, 0.5). We can define this value by experience.

To carry out an analysis of a resulting structure of links, it is appropriate to pass on from set of pairs of words with strong links to a weighted graph. There are words will be accepted as vertices, and the strong links between words will be accepted as edges. Obtained graph is too complex for visualization. One of the possible ways to attacking the complexity of this graph is to decompose it into smaller graphs (communities or components) (Fortunato, 2010). We use decomposition into strongly connected components. Between different strongly connected components of the graph there are no edges connecting them to each other. In an analyzed structure of links, strongly connected components correspond to sub themes existing inside the domain. We consider that these sub themes generate chains of associations, which correspond to strongly connected components.

5 Experiments

For carrying out of experiments, we use the Reuters-21578 Test Collection (Lewis, 1997). It is the most widely used collection of documents for text categorization and information retrieval tasks. Reuters-21578 consists of 21578 text documents, which belong to 135 categories (topics) mostly concerning business and economy.

For our experiments, we use only documents of 7 largest topics. Firstly, we apply the Chi-square test to allocate topic keywords. Table 4 shows the top of keywords of topic “Crude”. Further, for each topic we build co-occurrence matrix "words-by-documents" with the dimension of m X n (m is the number of key words and n is the number of documents of the training subset referred to the selected topic). Matrix cells contain frequencies of occurrence of words in documents. By applying singular value decomposition, we approximate the co-occurrence matrix by a matrix of a random rank k<=min(m,n). Then, using the cosine measure, we calculate the semantic distances between words in given row vectors in the new matrix. As it shown in examples in Table 3, the semantic links between words became more pronounced after SVD. At last, we build weighted graph for analyzing resulting structure of links. We use keywords as vertices. We connect two vertices (keywords) by edge if cosine distance between keywords surpasses 0.5. Since general graph is too complex for analysis, we decompose it into strong connected components.

Table 2: Top of keywords of topic “Crude”

No	Word	Chi2	No	Word	Chi2
1	oil	510,2584	11	prices	72,3330
2	crude	176,1830	12	gas	69,4536
3	barrels	166,8617	13	production	64,9400
4	petroleum	131,4903	14	exploration	61,9233
5	energy	118,6572	15	Texas	40,5523
6	day	111,3684	16	minister	40,4643
7	said	98,12144	17	industry	36,2952
8	barrel	97,33854	18	Venezuela	31,5506
9	bpd	90,42991	19	sea	30,8944
10	OPEC	78,40886	20	Gulf	30,3134

Table 3: Examples of discovered semantic links between words

Word	Close Word	Cosine distance	
		Before SVD	After SVD
OPEC	ceiling	0.6719075	0.7448371
	members	0.5199030	0.6264228
	opecs	-	0.6125585
	postponed	-	0.5974750
	quota	0.5227628	0.5821410
	output	0.5351974	0.5685460
	interview	-	0.5537803
barrel	sour	0.6884240	0,740011
	posted	0.7060034	0,735818
	intermediate	0.6620561	0,693202
	raised	0.6610617	0,68213
	price	0.6096797	0,644457
	Texas	0.6136986	0,642625
	Louisiana	0.5924909	0,621632
	brings	0.5808226	0,60325
	dlrs	0.5716912	0,602956
	increase	0.5205405	0,590747
	light	0.5651803	0,588729
	grades	0.5539949	0,586751
	sweet	0.5509997	0,57806
	prices	0.5034032	0,543379
	contract	-	0,511546
	crude	-	0,506415
	raise	-	0,505791
drill	area	-	0.5664646
	basin	-	0.5500753
	shelf	-	0.5121124

Figure 2 gives in detail one of the strong connected components of the graph corresponding to “Crude” topic.

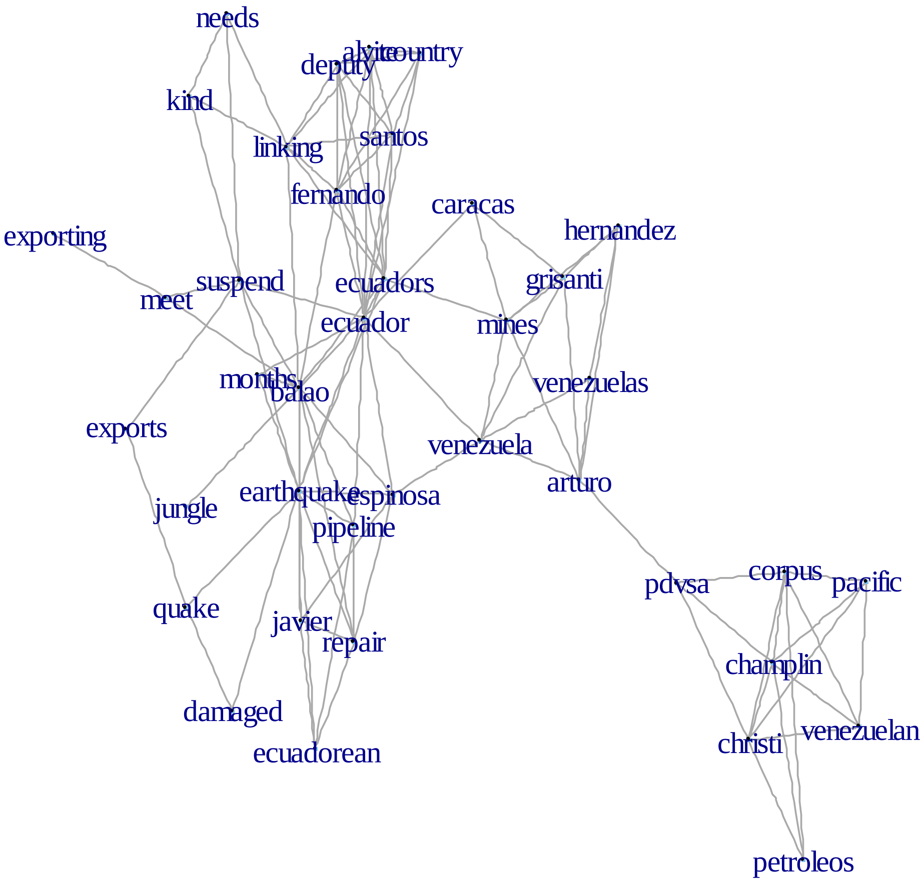

Figure 2: Subtopic “Ecuador” of topic “Crude”

6 Conclusion

Created thesaurus highly depends on source collection of documents. More representative collection of documents is required for reaching full reflection of domain terminology. However in this case dimension of graphs and matrices used in the given approach will require higher capacity computing power and probably development of processing algorithms on Map Reduce basis; further work will be related to this subject.

References

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000) A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. 21, 4, pp. 1253–1278.

Deerwester, S. C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A. (1990) Indexing By Latent Semantic Analysis. Journal of the American Society of Information Science, 41(6), pp. 391 – 407.

Ding, Y., & Engels, R. (2001) IR and AI: Using co-occurrence theory to generate lightweight ontologies. Proceedings of the 12th International Workshop on Database and Expert Systems Applications, pp. 961-965.

Dumais, S. (1990) Enhancing Performance in Latent Semantic Indexing. Behaviour Research Methods, Instruments, & Computers, 23(2), pp. 229-236.

Firth, J. R. (1957) A Synopsis of Linguistic Theory, 1930-1957. Special volume of the Philological Society. Oxford: Blackwell.

Fortunato, Santo. (2010) Community detection in graphs. Physics Reports 486.3, pp. 75-174.

Ido, D., Lee L., and Pereira, F. (1999) Similarity-based models of word co-occurrence probabilities. Machine Learning Vol. (34.1-3), pp. 43-69.

Ito, M., Nakayama K., Hara T., Nishio Sh. (2008) Association thesaurus construction methods based on link co-occurrence analysis for Wikipedia. Proceedings of the 17th ACM conference on Information and knowledge management, ACM, pp. 817-826.

Jing, Y., and Croft, W. B. (1994) An association thesaurus for information retrieval. Proceedings of RIAO, Vol. (94), pp. 146-160.

Khafajeh, H., Yousef, N., and Kanaan, G. (2010) Automatic query expansion for Arabic text retrieval based on association and similarity thesaurus. Proceedings of the European, Mediterranean & Middle Eastern Conference on Information Systems (EMCIS). Abu Dhabi, UAE, pp. 167-176.

Lancaster, Henry Oliver, and E. Seneta. (1969) Chi-Square Distribution. John Wiley & Sons, Ltd.

Lee, H. M., Lin, S. K., and Huang, C. W. (2007) Association Thesaurus Construction for Interactive Query Expansion Based on Association Rule Mining. Journal of Information Science and Engineering, Vol. (23.2), pp. 617-627.

Lewis, David D. (1997) Reuters-21578 text categorization test collection, distribution 1.0 [online]. Available from <http://www. research. att. com/~ lewis/reuters21578. html> [02.04.2014]

Lindén, K., and Piitulainen, J. (2004) Discovering synonyms and other related words. CompuTerm, pp. 63-70.

Manning. C.D., Raghavan P., and Schutze H. (2008) Introduction to Information Retrieval. Cambridge UP.

Morita, K., Arai, S., Kitagawa, H., Fuketa, M., and Aoe, J. I. (2011) Dynamic Construction of Hierarchical Thesaurus using Co-occurrence Information. Proceedings of the 2nd International Conference on Networking and Information Technology (ICNIT 2011), pp. 231-239.

Pinto, F.J., Martinez, A.F., and Perez-Sanjulian, C.F. (2008) Joining automatic query expansion based on thesaurus and word sense disambiguation using WordNet. International Journal of Computer Applications in Technology, Vol.(33), No.4, pp.271-279.

Schutze, H., and Pedersen, J.O. (1997) A Co-occurrence based thesaurus and two applications to Information Retrieval. International Journal of Information Processing and Management, Vol. (33), No 3, pp.307-318.

Sérasset, G., Chevallet, JP. (2004) Simple translations of monolingual queries expanded through an association thesaurus. Comparative Evaluation of Multilingual Information Access Systems. Springer Berlin Heidelberg, pp. 242-252.

Sinopalnikova, A. (2004) Word association thesaurus as a resource for building WordNet. 3rd Global WordNet Conference, Jeju Island, pp. 199-205.

Wang, P., Hu, J., Zeng, HJ., and Chen, Zh. (2008) Using Wikipedia knowledge to improve text classification. Knowledge and Information Systems, Vol. (19), No 3, pp. 265-281.

Zhang R., Sumita E. (2007) Boosting statistical machine translation by lemmatization and linear interpolation. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, pp. 181-184.