Web Page Classification Using WordNet's Linguistic Information

Eleftherios Kozanidis Vasiliki Simaki Athanasia Koumpouri

Department of Computer Engineering and Informatics, University of Patras, 26500 Patras, Greece

{kozanid, simaki, koumpour}@ceid.upatras.gr

Résumé Le World Wide Web fournit une grande quantité d'information dans plusieurs secteurs thématiques. En raison de sa nature dynamique, l'information que WWW nous offre s'augmente rapidement dans une base quotidienne. Par conséquence, une catégorisation du contenu de Web est très importante. La question de classification de pages Web est bien connue dans la communauté de la recherche d'information et d'apprentissage automatique, et soulève des questions importantes. Ainsi que les études dans le domaine de la classification matière-basée sont fortement développées, nous notons la manque de l'utilisation de l'information sémantique comme critère de catégorisation. Dans notre étude, nous essayons une catégorisation des pages Web dans des catégories thématiques prédéfinies, basées sur le traitement des textes d'ancrage des pages Web. Á ce but, nous exploitons l'information linguistique que les réseaux sémantiques nous fournissent pour l'anglais et la langue grecque. Nous notons également l'importance des traits de catégorisation évalués par des méthodes de désambiguisation sémantique.

Abstract The World Wide Web provides a large amount of information in several thematic areas. Due to its dynamic nature, the information that WWW offers increases rapidly on a daily basis. Therefore, a categorization of the web content assists Information Retrieval and Machine Learning purposes. The Web page classification issue is well known to the research community and raises important questions. Thus, researches in the topic-based Web page categorization are highly developed, we note lack in use of semantic information as a categorization criterion. In our study, we try a Web page categorization into predefined thematic categories, based on the processing of the Web pages' anchor texts. To this goal, we exploit the linguistic information that semantic networks provide us for the English and the Greek language. We note also the importance of the categorization features evaluated by semantic disambiguation methods.

Mots-clés : Classification des pages Web, Recherche d'information, WordNet , GreekWordNet, Méthodes de désambiguïsation sémantique, Mesures de similarité sémantique

Keywords: Web page classification, Information Retrieval, WordNet, GreekWordNet, Wold Sense Disambiguation methods, Semantic similarity measures

1 Introduction

Classification is a standard procedure which can be adapted in various fields. The totality of a database's content can be categorized into specific classes according to defined criteria. Considering the World Wide Web as a large database, a classification procedure of the web content seems necessary. The dynamic nature of the Web and the resulted continuous increase of information require as well a separate categorization of its total content. To this end, researches in the Web Page Categorization field try to replenish this need by proposing various types of classification. The Web Page classification is an important task for Information Retrieval and Machine Learning purposes. The informational plethora can be manageable and editable with categorization methods.

The Web Page classification resembles to the document classification process. However, Web pages are documents with no unified structure or consistency and the HTML language used offers to users, besides the text content, some additional information and other metadata. The Web page categorization procedure can be distinguished into different types, according to subject, function, sentiment or other criteria. In our study, we cope with the topic oriented categorization, namely in which thematic domain fits each selected Web page. The topic-based Web page categorization can be also separated according to the number of categories in which we try to assign a Web page.

In our study, we try to develop a classification system for Web pages in English and Greek, into predefined thematic areas that the Open Directory Project provides to users. The methodology we propose makes use the anchor text describing the selected Web pages. The anchor text constitutes the textual information that our system evaluates, in order to determine the thematic orientation of the Web page that describes and classify it to a certain category. This choice is important because the anchor text is a human-written shortcut of the thematic content of a Web page. Therefore, this is an acceptable, representative and reliable text sample of the describing Web page.

To develop our classification system we make use of the linguistic information that semantic networks provide us. We combine this information with Word sense disambiguation methods, in order to ameliorate the classification results.

2 Background work

The Web Page classification issue is the matching of a Web Page to one or more predefined categories. Qi and Davison (2009) in their work note the importance of Web-specific features and algorithms to categorize Web Pages and they describe state-of-the-art practices. In this important research, the authors proceed to the description of individual categories, based on the existed bibliography and they analyze the potential applications of a Web Page classification.

Wen & al. (2008) focus their research into a topic-oriented classification, in contrast to the functional categorization that Nanno et al. (2004) propose. Equally important are the Pang et al. (2002) and the Stamatatos et al. (2000) approaches, related to sentiment classification and genre categorization respectively. Mitchel (1997) in his survey deduces the categorization issue into a supervised learning problem, where a data set is delivered to the classifier in order to create the corresponding training model.

In our study we focus on thematic Web Page categorization, which can be made by using decision trees (Quilian, 1986, 1988), Bayesian categorization models (Cheeseman and Stutz, 1996) or hierarchical classifiers (Pulijala and Gauch, 2004, Sebastiani, 2002, Tie-Yan et al., 2011). Throughout our research, we

use the linguistic information needed to calculate the semantic similarity degree of senses that share an hyperonym (Turney, 2006, Turney and Pantel, 2010). We try also to resolve the ambiguity issue, which significantly affect the categorization, by proposing disambiguation techniques based on stored knowledge (Gonzalo et al., 1998).

3 Semantic Networks and Word Sense Disambiguation methods

3.1 Semantic Networks

A semantic network is a graph structure, often used as a form of knowledge representation. It consists of nodes and edges. The nodes represent concepts and the edges demonstrate the semantic relations between concepts. The semantic networks share several features with traditional dictionaries of synonyms and thesaurus. They provide information not only for every entry's sense but also for the entries that are semantically related. In other words, a semantic network is an electronic lexicographic resource stored in lexical databases, which organize the semantic relationships of words in a hierarchical structure.

In our study we used WordNet (Felbaum, 1998, Miller, 1990) which is the lexical database of English. Unlike preexisting dictionaries, the WordNet entries are grouped into sets of cognitive synonyms (synsets) which express distinct concepts. The synsets are linked to concepts that represent semantic and lexical relations, and they form a semantic network. Due to this structure, WordNet has been considered by NLP research community as a useful tool for computational linguistics and natural language processing.

The Greek WordNet linguistic resource was developed under the research program BalkaNet[1] (Tufis et al., 2004, Stamou S. et al., 2002, Oflazer et al., 2001) aiming the development of a multilingual semantic lexicon according to the principles of EuroWordNet. The languages included in BalkaNet are: Bulgarian, Czech, Greek, Romanian, Serbian, and Turkish. The Greek Wordnet has been implemented by the standards of the BalkaNet resource. The followed practices are mainly related to the representation of common concepts among all languages involved. An example of this correlation is the connection way of all monolingual synonym sets to their respective concepts of WordNet 2.0 and the adaptation of the English WordNet semantic relations such as hyponymy, synonymy, and meronymy relations, which are required for the WordNets of all BalkaNet languages. Beyond the common terms for all Balkan languages, the Greek WordNet has been enriched with additional concepts, in order to meet the requirements of a conventional dictionary. Furthermore, an additional electronic lexical resource has been performed rather than retrieving the terminology of every word.

3.2 Word Sense Disambiguation Methods

Disambiguation is the process of identifying the most appropriate sense for every text's word, from a set of alternative senses which can represent a word. It is directly related to the problem of polysemy, an inherent characteristic of natural languages' words, whereby it is possible for a word to be assigned to more than one sense. The IR society has proposed a number of different techniques that can be used to resolve the ambiguity in a text. In our study, we used techniques based on stored knowledge in a dictionary, thesaurus or knowledge databases without using text collections. Our methodology is one of the so-called dictionary-based methods, and during our work we used the WordNet semantic network as the knowledge database.

[1] http://www.ceid.upatras.gr/BalkaNet/

The term *semantic similarity* refers to the semantic relationship between two concepts that share an hypernym. Two words are semantically related to each other when they display any kind of semantic relation, while the more properties they share the closer they are semantically (Turney, 2006). In our study, we calculate the semantic similarity by using two different categories of semantic similarity measures. The first category, called e*dge counting* includes measures which calculate the semantic similarity between concepts by considering the edge distance separating sets of synonyms that express the corresponding concepts in the concept hierarchy. The edge counting measures we studied are: wup (Wu and Palmer, 1994), lch (Leacock and Chodorow, 1998), and li (Li et al., 2003). The second category is called *Information Content* and includes measures considering the principle that the similarity of two concepts is associated to the shared information. The information content is defined for each concept separately expressing how specific is the information that a concept carries. The information content value of a concept is defined by the concept presence possibility in a large text collection. The Information Content measures that we studied are: res (Resnik, 1995), jnc (Jiang and Conrath, 1997) and lin (Lin, 1998).

The disambiguation algorithm used in our study is a variation of the UMND1 algorithm (Patwardhan et al., 2007). The proposed algorithm takes as an input a list of terms that do not contain duplicate records. The list resulted after the morphosyntactic annotation (PoS Tagging) and the lemmatization of the input text, consists of nouns or collocations included in WordNet noun database. We decided to use only nouns because our methodology uses the hypernyms/ hyponyms hierarchy that WordNet provides only for nouns and verbs. The same methodology can be applied to verbs, but in this study we focused on nouns' disambiguation based on the fact that nouns carry the text's thematic information.

The algorithm disambiguates one word at a time, using a frame of n terms before and after the word of interest. Given a term t that we attempt to disambiguate, let T be the set of different senses and let $t_1, t_2,\ldots t_t$ be each of these senses. Let $w_1,w_2,\ldots w_{n-1}$ be the terms of the extract and $W_1, W_2,\ldots W_{n-1}$ the corresponding different senses of each of them. For each of the candidate senses t_i of the term t, we calculate a score using the following formula:

$$score(t_i)=\sum_{j=1}^{n-1} \max_{k=1 \text{ for } w_j} (similarity(t_i, w_{jk}))$$

where w_{jk} is the k-th sense corresponding to the w_j term. As similarity measure we used those presented in previous section. The sense chosen for each term t_i is one that achieves the highest *score(t_i)*. In case that two senses of t_i term achieve the same *score(t_i)* we choose the sense which, according to WordNet, represents the concept of t_i more often.

3.3 Results

In order to evaluate our disambiguation methodology, we used SENSEVAL-3 data (Snyder and Palmer, 2004). Our evaluation data is a small subset of «Penn Treebank» collection and consists of three articles. The first two of them articles are from «Wall Street Journal» newspaper and the third is an extract from «Brown» corpus. The extracts are divided into sentences and the words included are disambiguated and noted according to WordNet's 1.7 concept codes by linguists. More particularly, we used SENSEVAL-3

English all-words[2] control data that were properly formatted in order to follow the SemCor formalism and we modified the synsets encodings to comply with WordNet 2.0.

In each performance, we applied a different measure to calculate the semantic similarity between the candidate concepts for each pair of words. We proceed to count the following three evaluation indicators: the noun disambiguation accuracy (1), the polysemous noun disambiguation accuracy (2) and the disambiguation accuracy of polysemous nouns in phrases with more than one nouns (3). The semantic similarity methods applied are: Resnik, JiangConrath, Lin, LeacockChodorow, WuPalmer, Li.

	FirstSense	*Random*	*Jiang Conrath*	*WuPalmer*	*Resnik*	*LeacockChodorow*	*Lin*	*Li*
1	67%	36,56%	47,92%	43,96%	41,72%	42,52%	47,44%	43,64%
2	61,72%	23,60%	37,80%	32,88%	30,08%	31,08%	37,24%	32,48%
3	61,24%	23,56%	36,04%	30,84%	27,88%	28,92%	35,44%	30,40%

Table 1 : The semantic similarity results for each of the three cases.

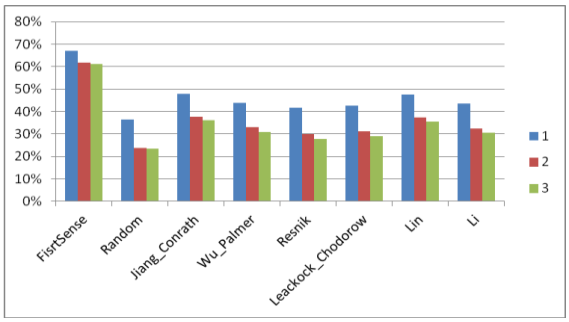

Figure 1: The results of disambiguation accuracy.

Our experimental results indicate that our proposed technique achieved sufficient accuracy in the disambiguation process. Furthermore, we noticed that best results were achieved by using the JiangConrath similarity measure (47.92% accuracy), and followed by Lin measure (47,44% accuracy). Moreover, the time complexity of our algorithm allows us to disambiguate texts in real time and makes it possible to incorporate it into focused crawling applications. The performance of the proposed system was evaluated for the English language due to the lack of evaluation data for the Greek language. However, the application can be also used to disambiguate text written in Greek.

4 Web Page Classification

4.1 Topic-based Categorization

In this chapter we present a new approach of Web Page classification, based on the semantic information evaluation contained in the anchor texts of the links. Even though the manual review of a Web page's content and its matching into thematic classes by human web editors is more reliable, the dynamic Web nature and the large volume of information leads us to an automated categorization system. The most

[2] http://www.cse.unt.edu/~rada/downloads.html#sensevalsemcor

ambitious attempts to this direction are the Yahoo! thematic Internet lists[3], the Open Directory Project[4], the Google Directory[5] and the Looksmart[6]. Through a predefined hierarchy of categories, the user is able to access web pages of a specific topic that responds to his current informational needs.

4.2 Thematic classifier: Implementation and training

Our classifier can distinguish the thematic content of a Web page by taking into account the anchor text of its link. The classifier decides in which categories a website can be assigned by choosing among a set of 245 categories for the English language and a set of 13 categories for the Greek language. The difference in the number of classes is due to the lack of available examples of classification for the Greek language. For the definition of categories for the English language, we select 13 categories (News, Science, Society, Sports, etc) of second level of the Open Directory Project (ODP) hierarchy. For the Greek language, we select the corresponding categories that are located under the category of Top-World-Greek.

To categorize the pages in the English language, we extended further the set of available categories by replacing each of the 13 second-level categories with the corresponding third-level categories. Overall we chose 245 categories. For each category we selected about 500 examples of Web pages that Open Directory Project authors have manually categorized under this category. Each Web page instance should be assigned exclusively to one category, which is why we remove the training examples that are mapped in more than one category. For every instance, ODP provides a sentence that describes the content of the Web page and emulates the anchor text that describes the Web page's content in the Web.

Since the text that represents a Web page is defined, we attempt to set the representative features for each classification category. As a categorization characteristic we define the textual feature extracted from a representative Web page text belonging to a distinct category. As a textual feature we consider a word or the code of the set of synonyms in which the word is assigned after the disambiguation process (synset Id), the combination of the word to the corresponding synsetid, or the combination of two similar of these features by two. We choose to examine word pairs according to Turney's principle (Turney, 2008), whereby the word pairs that co-appear in similar environments tent to have semantic relations.

We used words that represent either nouns or proper nouns, as these terms are considered representative of the thematic orientation of a text. Finally, in cases where we used as a categorization characteristic synsetids resulting from the disambiguation process of corresponding terms, we experimented with many different disambiguation methods. To define categorization features, we processed the text that describes each web-example as follows. Firstly, we comment morphosyntactically the text using TreeTagger when it comes to text written in English and GreekTagger when text is written in Greek. Then, we trace the collocations based on WordNet term index. If two consecutive terms appear in WordNet index as one, then the term is recognized as collocation. The terms resulted from this process should belong to the set of nouns as to the part of speech, otherwise they are discharged.

---

[3] http://www.yahoo.com/

[4] http://www.dmoz.org/

[5] http://dir.google.com

[6] http://search.looksmart.com

Since we defined the categorization feature of interest, we set the importance of this feature for a category. To this end, we apply a variant of TF-IDF method (Salton and McGill, 1983, Salton and Buckley, 1988) in collecting categories defining as document the text derived by bringing together representative texts of all pages belonging to one category. Specifically, to calculate the weight of a feature term t for a category C we based on the variation of TF-IDF measure (Wang et al., 2010).This process demonstrates how representative is a classification term for one category over others. The logic of the measure is summarized in the following statement: in fewer categories a term appears, the more strength it gains as a separator. The results of our experiments are listed in the following section.

4.3 Experimental results and discussion

To evaluate the effectiveness of our classifier, we divided the set of websites that we have assigned to each category in 10 subsets with each containing the 1/10 of the total number of initial set (50 Web pages). Then we use each time 9 of the 10 subsets as training data in order to extract weights for each classification term and create the corresponding indexes. The rest of the data set was not used as testing set. We implemented a set of classifiers and in table 2 we present the obtained results for the English.

Categorization feature	***Semantic similariry measure***	***AVG P***	***AVG R***	***AVG fs***
noun		0,50465	0,441978	0,417687
WordNet term		0,51139	0,450489	0,429267
disambiguated terms	WuPalmer	0,435779	0,369486	0,367193
	JiangConrath	0,443192	0,340721	0,353042
	Lin	0,413486	0,34817	0,345284
	First Sense	0,398161	0,337423	0,321513
Co-appearing synsets	WuPalmer	0,710542	0,627712	0,631778
	Li	0,687303	0,591391	0,599539
	LeackockChodorow	0,744568	0,680101	0,68034
WordNet sysets		0,787084	0,745673	0,742593
Co-appearing WordNet sysets, semantically related	JiangConrath	0,758438	0,698909	0,698857
	LeackockChodorow	0,754471	0,692072	0,693094
Co-appearing Noun pairs		0,765857	0,72937	0,720482
synsets and	Lin	0,536532	0,467531	0,449233

	JiangConrath	0,527779	0,463659	0,44309
	WuPalmer	0,529092	0,468488	0,446161

Table 2 : The classification results for the English.

Based on the results of experimental evaluation of our classifier, we found that the proposed classification system decides properly, graduating the proper category as first between 246 categories, at least in half of the experiment cases. The classifier's performance ameliorates appreciably if loosen the classification criterion of success, assuming that the classifier's choice is correct if the requested class is among the first 5 of 246. An indicative example: when a co-appearing WordNet synset, semantically related is considered as the categorization feature, the average accuracy increases from 0.75 to 0.91, the average recall from 0.69 increased to 0.88 and average fscore from 0.69 increased to 0.88. We observe, though, the need of improving the disambiguation system whose performance is marginally acceptable and seems to be responsible for the unsatisfactory classifier performance in cases where only one of the WordNet synset term is disambiguated. These observations are consistent to Gonzalo et al. observations (Gonzalo et al., 1998) whereby the performance of an information retrieval system ameliorates when disambiguated terms are used as synsets, as the retrieval system's efficiency for a rate of incorrect disambiguation up to 30% is better than a conventional system, while the retrieval system's efficiency for a rate of incorrect disambiguation of 30% to 60% is at least similar with a conventional system.

Categorization feature	***Semantic similariry measure***	***AVG P***	***AVG R***	***AVG fs***
Noun		0,711983	0,701522	0,694092
WordNet term		0,669499	0,663844	0,652132
Disambiguated terms	Wu Palmer	0,687236	0,694255	0,677686
	Jiang Conrath	0,690783	0,695882	0,682488
	Resnik	0,683278	0,694813	0,67422
	Leackock_Chodorow	0,688513	0,697673	0,682078
	Lin	0,680914	0,696788	0,675982
	Li	0,694061	0,70367	0,684404
	First Sense	0,609767	0,595765	0,581916
synset	Wu Palmer	0,502615	0,362829	0,413949
	Jiang Conrath	0,484208	0,3521	0,400448
	Leackock_Chodorow	0,484064	0,347049	0,396591
Co-appearing WordNet synsets		0,872439	0,745927	0,80218

Co-appearing WordNet sysets, semantically related	Jiang Conrath	0,810651	0,674241	0,733072
	Lin	0,754785	0,47326	0,577827
Co-appearing Noun pairs		0,900826	0,809864	0,851244
synsets and disambiguated synsetId	Lin	0,692954	0,708576	0,688033
	Jiang conrath	0,700065	0,704761	0,69153
	Wu Palmer	0,703379	0,70703	0,692468
SVM : WN terms		0,789	0,743	0,738

Table 3 : The classification results for the Greek.

Given the results of this experiment, we conclude that the categorization results for the Greek language, although within acceptable levels are not improved by the use of the semantic network. We also note that the disambiguation methods using semantic similarity measures show improved performance in classification than the First Sense method which certifies their properness to disambiguate nouns.

The reasons for the low performance of our classification system can be found on both the small size of the Greek WordNet (only 24366 terms), and the disambiguation system performance. Results could be improved by enriching the training examples which in this case can be described as relatively poor (only 2360 pages). We should also mention that the classification criterion of success, according to which a Web page should be classified first among 13 candidate categories in order to be regarded as a successful, can be considered relatively strict.

Finally, we compared the results of our experiment with the classification results we got by training a Support Vector Machine classifier (SVM) provided through the package WEKA[7], using the same data. From the results we found that the use of classifiers improves the classification performance when WordNet nouns are used as categorization features. Nevertheless, the performance of the proposed technique approaches the performance of SVM classifiers when word pairs and disambiguated synsetId were used as categorization feature.

5 Conclusions and Future Work

In our study, we evaluated the utility of semantic networks in Web page classification. In particular, we suggested and implemented algorithm designed to assess both WordNet and GreekWordNet as resources for a topic-based Web page classification. Our experimental results indicate the effectiveness of the proposed classification system to identify the correct category applicable to a Web page.

As future work we aim to improve the efficiency of the disambiguation method which is the most likely cause for the cases of reduced performance of our classification system. An interesting approach would be the use of adjectives that come along with nouns in order to disambiguate the latter's term sense. In terms of improving the performance of the thematic categorization technique we implemented, a challenging

[7] http://cs.waikato.ac.nz/ml/weka

research direction is to combine our technique to classification features derived from links Web pages of interest. We consider the contribution of semantic networks and the techniques developed, as the cornerstone in the development of focused crawling applications and we hope that research conducted in this study will inspire future studies.

References

Cheeseman, P., Stutz, J. (1996): "Bayessian Classification (Ayto Class): Theroty and Results" Advances in Knowledge Discovery and Data Mining (Eds: U. Fayyad et al.) AAAI Press.

Fellbaum Ch. (ed.) (1998): "*Wordnet: An Electronic Lexical Database"*. MIT Press.

Jiang j. & Conrath. D. (1997): "Semantic similarity based on corpus statistics and lexical taxonomy". Proceedings of the International Conference on Research in Computational Linguistics, Taipei, Taiwan.

Leacock. C., and Chodorow. M. (1998): "Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet". In C. Fellbaum, editor, An Electronic Lexical Database, pages 265–283. MIT Press, 1998.

Li Y., Bandar Z. A., and McLean D. (2003): "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources". IEEE Trans. On Knowledge and Data Engineering, 15(4):871–882, July/Aug.

Lin D. (1998): "An information theoretic definition of similarity", Proceedings of the International Conference on Machine Learning, Madison, U.S.A., 296-304

Miller G. A. (1998): "Nouns in WordNet". In *WordNet: An Electronic Lexical Database,* C. Fellbaum (ed.) pp. 23-46.

Mitchell, T. M (1997): "Machine Learning". McGraw-Hill, New York, NY

Nanno, T., Fujiki, T., Suzuki, Y., and Okumura, M. (2004): "Automatically collecting, monitoring and mining Japanese Weblogs". In proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters (WWW Alt.). ACM Press, New York, NY, 320-321

Oflazer K., Stamou S., Christodoulakis D. (2001): "BALKANET: A Multilingual Semantic Network for the Balkan Languages". In the *Elsnet Newsletter,* vol. November 2001.

Pang ,Bo., Lee, L., and Vaithyanathan Sh.(2002): "Thump up? Sentiment Classification using Machine Learning Techniques" In Proceedings of th Conference on Emprical Mathods in Natural Language Processing (EMNLP) pp 79-76

Patwardhan S., and Pedersen T., (2006): "Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts". In proceedings of the EACL 2006 WorkShop on Making Sense of Sense: Bringing Computational Linquistics and Psycholinguistics Together, pages 1-8, Trento, Italy, April.

Pulijala A., Gauch S. Hierarchical Text Classification. International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2004 Orlando, FL, July 21 - 25, 2004

Qi Xiaoguang and Davison Brian (2009): “Web Page Classification: Features and Algorithms”. In ACM Computing Surveys, Vol 41, No 2, Article 12 (Feb 2009)

Quilian J. R. (1988): “Decision Tress and Multi-Valued Attributes”. In *Machine Intelligence* 1: 305-318.

Quilian J.R. (1986): “Induction of Decision Trees”. In *Machine Learning* 1(1):81-106.

Resnik. P.,(1995): “Using information content to evaluate semantic similarity”. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, 448 453

Salton G., Buckley C. (1998): “Term-weighting approaches in automatic retrieval”. Information Processing & Management, 24(5): 513-523.

Salton G., McGill M. (1983): *Introduction to Modern Information Retrieval.* Singapore: McGraw-Hill.

Sebastiani, F. (2002): “Machine learning in automated text categorization”. ACM Computing Surveys, 34(1), 1-47

Snyder, B., Palmer, M. (2004): “The english all-words task”. In: Proc. of Senseval-3, pp. 41–43

Stamatatos, E, Fakotakis, N., and Kokkinakis, G (2000): “Automatic text categorization in terms of genre and author”. Computational Linguistics, 26(4), 471-495

Stamou S. Oflazer K., Pala K., Christodoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D. Grigoriadou M. (2002): “BALKANET: A Multilingual Semantic Network for Balkan Languages”. In *Proceedings of the 1st International Global WordNet Conference* (GWC), Mysore, India, Jan. 21-25 2002.

Tie-Yan Liu,Yiming Tang, Hao Wan, Hua-Jun Zeng, Zeng Chen, and Wei-Ying Ma, “Support Vector Machines Classification with A very Large-scale Taxonomy” in Proceedings of CICLING 2011 Tokyo, Japan (2011).

Tufis D., Cristea D., Stamou S. (2004): “Balkanet: Aims, methods, results and perspectives. A general overview”. *Romanian J. Sci. Tech. Inform. (Special Issue on Balkanet),* 7(1-2), pp. 9–43.

Turney, P. D. (2008). “The latent relation mapping engine: Algorithm and experiments”. Journal of Artificial Intelligence Research, 33, 615-655.

Turney, P. D. 2006. “Similarity of semantic relations. Computational Linguistics”, 32 (3),379{416.

Turney, P., & Pantel, P. (2010): “From frequency to meaning: Vector space models of semantics”. Artificial Intelligence Research, 37, 141–188.Gonzalo J., Verdejo F., Chugur I.and Cigarran J., (1998): “Indexing with WordNet synset can improve text retrival”. In Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP, Montreal, 1998

Wang D., Zhang H., Wu W.,& Lin M., (2010): “Inverse category frequency based supervised term weighting scheme for text categorization”

Wen Hao., Fang Liping., & GuanLing (2008), “Automatic Web Page Classification Using Various Features”, PCM 2008 LNCS 5353, pp 368-376 Berlin Heidelberg

Wu Z.and Palmer M. (1994): “Verb Semantics and Lexical Selection”. In *Annual Meeting of the Associations for Computational Linguistics (ACL’94)*, pages 133–138, Las Cruces, New Mexico, 1994