Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

DOSCOS Engine: An Automated Approach to Collect Domain-Specific Corpora through Search Engines

Santhi Priya Polisetty[*]
Research scholar, CSE Dept,
University College of Engg & Tech,
Acharya Nagarjuna University,
Nagarjuna Nagar, Guntur, AP, INDIA.

T.V .Rao
Department of CSE,
KL University (KLU),
Vaddeswaram, Guntur,
AP, INDIA

Abstract— ***Social networking today has become a very popular communication tool among Internet users. Millions of users of these websites share opinions on different aspects of life every day. Opinions are important because whenever a person needs to make a decision, he wants to hear others' opinions. So, social networking websites are rich sources of data for opinion mining and sentiment analysis .Every social networking site has it's own search engine. In this paper we propose a methodology to collect domain-specific corpora from the Internet, using a search engine based approach, an alternative to using news agencies or focused crawling which will supposedly obtain more varied corpora .The method we propose for obtaining specialized corpora on a language is based on the BootCaT method. But , instead of the seed words, a sample mini-corpus is used as a basis for the process . In this paper we present the research we have done to build a large general corpus of pollution freedom from the web. We have tried and evaluated by using our proposed search engine, DOSCOS engine and by BootCat tool, in terms of parameters such as speed, cost, size or quality. The results show that our results are comparable to or even better than previous findings.***

Keywords—***social network ,domain-specific corpora , search engine , document , seed word , query , filter***

I. INTRODUCTION

The Social Network Analysis deals with the analysis of the relationships that exist between entities in a social network. For instance, in a social network of people, the analysis can include who is friend with whom, who can influence which group of people, whom can have access to the information that goes through this network etc. Lately there has been a growing interest in this field, especially as to how it gets involved with knowledge discovery and data/web mining. For instance, analysing the behaviour of users in online discussions or discover how users form communities and are affected by them are interesting works. Language usage is dependent on domain (Hanks, 2000) and domain specific corpora are consequently extremely useful for language learning and lexicography (Barrière, 2009; Drouin, 2004). It is possible to label heterogeneous data for domain either manually (Atkins et al., 2010) or automatically (for a survey see (Sebastiani, 2002)) using human knowledge or machine learning. State-of-the-art text classification uses supervised techniques whereby a system learns from previously annotated data. This works well when such data is available in sufficient quantities for supervised machine learning, though often that is not the case depending on the domain and language required. Moreover, this approach assumes that the heterogeneous data in the available corpus covers the required domains.

Fisher et al. (2006) analyse newsgroups by applying Social Network techniques and they interpret online communities by assigning roles to the members of the groups. This is done by observing how people relate to each other in a graph-based model of post-reply relations. They notice that short discussion threads point out question-answer exchanges and longer threads indicate proper discussions. Java et al. (2007) analyse the Twitter's social network and the intentions of the associated users in order to understand the reason why people use such networks. They identify the communities that are formed, they categorize them into communities that create information, communities that receive information and communities that exist only because of friendship. They label the identified communities by the keywords that appear in the various posts. In early studies, association coefficients were generally used as association measures in term or document clustering.Among these, cosine coefficient and Jaccard coefficient are most frequently used in measuring term association(Augustson & Minker, 1970; Dillon & Caplan, 1979;Grefenstette, 1994; Kim & Choi, 1999; Salton & Lesk,1971; Yu & Raghavan, 1977). Mutual information, a measure based on Shannon's information theory, was applied in an analysis of lexical collocation (Church & Hanks, 1990) and has been used in many applications thereafter (Gauch & Wang, 1997; Johansson, 1996; Kang & Choi, 1997; Kim & Choi, 1999). Recently, the measures based on test statistics such as x2 and *z*-score and likelihood ratio have been proposed as alternatives to the earlier association measures (Delcourt, 1992; Dunning, 1993; Kageura, 1997; Plaunt & Norgard, 1998).

Although there are numerous association measures tested in many applications, little research focusing on comparing the performance of the measures in a single experiment has been done. In a comprehensive study evaluating factors affecting document ranking, McGill et al. (1979) reviewed 67 similarity measures, and found that there is a significant difference in the performance of association measures. Recently, Kageura (1997) tested x^2, likelihood ratio, Yule's coefficient of colligation *Y*, and mutual information. Among the four measures examined in his bigram extraction experiment, x^2 and likelihood ratio generally performed better than Yule's *Y* and mutual information. In particular, the likelihood ratio performed best when the size of corpus was relatively small, whereas x^2 did better as the size of corpus grew. Kageura claimed that likelihood ratio is the most suitable for his application. In a query expansion experiment, Kim and Choi (1999) compared five association measures including three similarity coefficients, average conditional probability, and normalized mutual information. In their study, the three similarity coefficients, namely, Jaccard's, Dice's, and cosine, performed better than the other measures did.

The development of Web2.0 has resulted in the generation of a vast amount of blog repositories, review sites, web forums and online discussions. In this type of discussions people express opinions, criticize products and ideas, exchange knowledge and beliefs. Tracking opinions on specific subjects allows the identification of user expectations and needs, feelings of people about certain political decisions or reactions against particular events.

The method we propose for obtaining specialized corpora on a language is based on the BootCaT method (querying search engines for random combinations of a list of seed words representative of the domain or topic and retrieving the pages returned) but, instead of the seed words, a sample mini-corpus is used as a basis for the process: most representative words are automatically extracted from it .The approach most closely related to ours is that used by the BootCaT tool (Baroni and Bernardini, 2004), which introduced a new methodology for collecting monolingual domain-specific corpora from the Internet: give a list of words as input, query APIs of search engines for random combinations of these seed words and download the pages. This methodology has in some cases been used to build big general corpora (Sharoff, 2006), but for collecting smaller specialized corpora, it has become the de facto standard, replacing focused crawling. Although BootCaT is a monolingual corpora collector, we can expect that, by applying it to word lists on the same subject but in different languages, one could obtain light multilingual comparable corpora.

II. Our Approach

The aim of our research project is to develop a methodology to collect domain-specific comparable corpora from the Internet, using a search engine based approach similar to that of BootCaT. The few studies that the authors have found on the topic precision obtained by BootCaT's word-list method show that this is not at all perfect (Baroni and Bernardini, 2004; Leturia et al., 2008a). Thus, improving the domain-precision of the corpora obtained is one of our goals. In order to try to improve the domain-precision in monolingual corpus collection of the BootCaT method, our approach takes a sample mini-corpus of documents on the topic, instead of a list of words. The list of keywords to be used in the queries is automatically extracted from the mini-corpus. Mini-corpus is also used to filter the downloaded documents according to the domain by using document-similarity techniques (Lee et al., 2005). Next we will describe the whole process we use for obtaining monolingual domain-specific corpora, which is inspired by Leturia et al. (2008a), step by step and in more detail:

A. *Collection of sample mini-corpus manually:*

The sample mini-corpus of documents on the target domain, which is the basis of our system, has to be collected manually. The criteria when collecting the sample is that it should be as heterogeneous as possible and cover as many different subjects of the domain as possible.

B. *Extraction of keywords automatically :*

The seed words to be used in the queries are extracted automatically from the sample corpus, with the TF-IDF method .A keyword or topic for a document is a word or multi-word (sequence of 2 or more words) that summarizes in itself part of that document content. Tf-Idf

metric. Tf-Idf (Term frequency-Inverse document

frequency) (1) is a statistical metric often used in information retrieval and text mining. Usually, it is used to evaluate how important a word is to a document in a corpus. The importance increases proportionally to the number of times a prefix/word/multiword appears in the document but it is offset by its frequency in the corpus. It should be noticed that we use a probability, p(W, dj), in equation (1), defined in equation (2), instead of using the usual term frequency factor.

$$\text{Tf-Idf}(W, dj) = p(W, dj) * \text{Idf}(W, dj). \qquad (1)$$

$$p(W, dj) = f(W, dj) / Ndj . \qquad (2)$$

$$\text{Idf}(W, dj) = \log(\| D\| / \| \{ dj : W \in dj \}\|). \qquad (3)$$

Where f(W,dj) denotes the frequency of prefix/word/multiword W in document dj and Ndj stands for the number of words of dj; || D|| is the number of documents of the corpus. So, Tf-Idf(W,dj) will give a measure of the importance of W within the particular document dj . By the structure of term Idf we can see that it privileges prefixes, multi-words and single words occurring in fewer documents.

C. *Part-of speech tagging:*

The mini-corpus is POS-tagged, and then the most significant nouns, proper nouns, adjectives, verbs, entities are extracted using CST's Part-Of-Speech tagger ,Brill, with adaptations(which is available online for free) The POS-tagger marks each word in a text with information about word class and morphological features, for example "This page is about the Brill-tagger" →

This/DT page/NN is/VBZ about/IN the/DT Brill-tagger/NNP

In order to maximize the performance of the queries, the extracted list can be revised manually, to remove too specific or too local proper nouns, words that are too general and polysemous words that have other meanings in other areas.

D. Querying search engines and downloading:

Random combinations of the extracted seed words are sent to the APIs of search engines and the pages returned are downloaded, just as in the BootCaT method. If performance of search engines is very poor, mostly due to the rich morphology of the language ,then solve this problem by means of morphological query expansion, which consists of querying for different word forms of the lemma, obtained by morphological generation, within an OR operator.

E. Language filter:

For filtering content that is not in the target language out of bilingual documents, we use LingPipe, A tool kit for processing text using computational linguistics. LingPipe's text classifiers learn by example. For each language being classified, a sample of text is used as training data. LingPipe learns the distribution of characters per language using character language models. Character language models provide state-of-the-art accuracy for text classification. Character-level models are particularly well-suited to language ID because they do not require tokenized input; tokenizers are often language-specific.

F. Length filter:

Filtering documents by length is an effective way of reducing noise (Fletcher, 2004). In our case, we reject documents the length of which after conversion to plain text is under 1,000 characters or over 10,000 characters.

G. Near-duplicates detection:

We have also included a near duplicate detection module based on Broder’s shingling and fingerprinting algorithm (Broder, 2000).

H. Domain filtering:

We represent both the downloaded documents and each of the documents of the sample corpus with a vector of the most significant keywords, i.e. nouns, proper nouns, adjectives and verbs. These were extracted through a POStagger. The keywords are selected and weighed by some frequency measure, such as LogLikelihood Ratio .

I. Similarity measure:

For measuring the similarity we use the cosine, one of the most widely used ways to measure the similarity between documents represented in the vector space model. A document is accepted in the corpus if the maximum of its cosine measures with each of the documents in the sample mini-corpus reaches an empirically defined threshold, and rejected otherwise.

We depicted the DOSCOS engine process as a flow graph in fig.1

III. Evaluation

we collected the sample mini-corpora used for pollution freedom of 22 short articles (about 4,000 words) obtained from popular natural science magazine. We experimented the pollution freedom mini-corpora on both the two methods ,the traditional method BootCat method and our proposed method DOSCOS engines. In order to see which of the two methods obtains a higher degree of comparability we collected two english comparable corpora with each of the two methods mentioned above.

Let the basic contingency table is

TABLE 1 contingency table for term co-occurrences

		Term Y		
		Present	**Absent**	
Term X	**Present**	*a*	*b*	*a+b*
	Absent	*c*	*d*	*c+d*
		a+c	*b+d*	*n*

Then, for evaluating the two methods, we used the measure the comparability of the two corpora obtained: by calculating Mutual Information (MI) statistic (Church and Hanks 1989).

$$\mathrm{MI}_{w,X} = \log_2((a/a+c) * (N/a+c))$$

MI ,the (log of the) ratio of the word’s relative frequency in one corpus to its relative frequency in the joint corpus. This is an information theoretic measure

(with relative frequencies serving as maximum likelihood estimators for probabilities) as distinct from one based in statistical hypothesis testing, and it makes no reference to hypotheses. Rather, it states how much information word *w* provides about corpus *X* (with respect to the joint corpus). It was introduced into language engineering as a measure for co-occurrence, where it specifies the information one word supplies about another.

Church and Hanks state that MI is invalid for low counts, suggesting a threshold of 5. In contrast to χ^2, there is no notion in MI of evidence accumulating. MI, for our purposes, is a relation between two corpora and a word: if the corpora are held constant, it is usually rare words which give the highest MI. This contrasts with common words tending to have the highest χ^2 scores. Church and Hanks proposed MI as a tool to help lexicographers isolate salient co-occurring terms. Several years on, it is evident that MI overemphasises rare terms, relative to lexicographers' judgements of salience, while $\chi^{□2}$ correspondingly overemphasises common terms. In terms of Table 2 we depicted experimental results

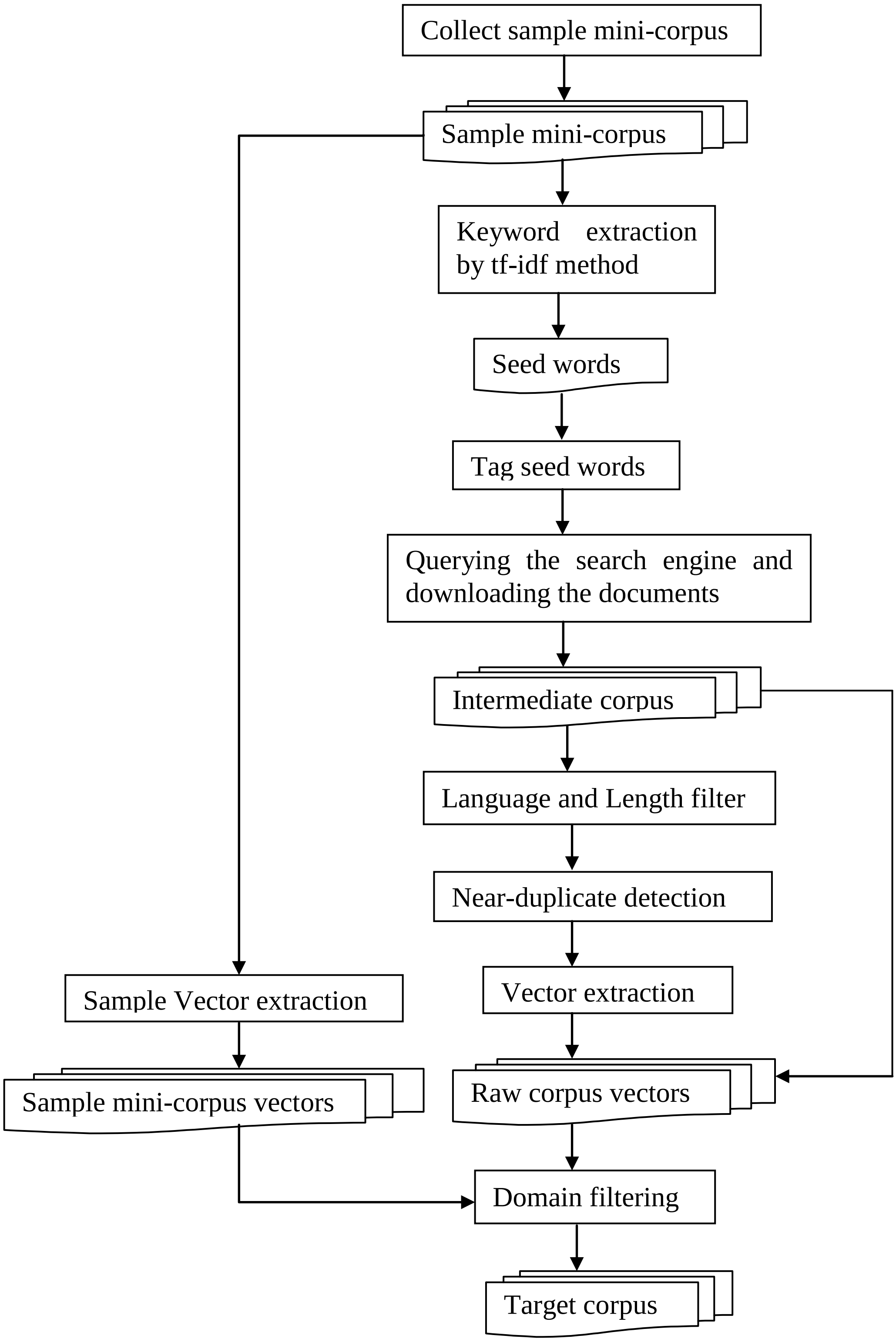

Fig. 1 A flow graph for DOSCOS engine

Table 2. Evaluation results

Corpus	**Method**	**Mutual Information statistic(MI) for n keywords**			
		500	**1000**	**5000**	**10000**
Pollution Freedom	BootCat and DOSCOS engine	0.064	0.314	0.42	0.50

IV. CONCLUSION

This paper has presented a search engine based method for collecting, high domain-precision, monolingual corpora out of sample mini-corpora corpora from the Internet. We tried a variant of BootCat method that uses only one sample mini-corpus to see if we could obtain similar or better comparability with less initial effort .However, this evaluation cannot be considered conclusive, for the following reasons:

1. The evaluation was done with only one corpora. Besides, we now believe that pollution freedom might not have been a good domain choice for the evaluation, because it does not completely fit into what we know as a specialized domain (interdisciplinary terminology, etc.).Evaluations with more corpora and more domains are needed before stating anything definite.
2. There is not much literature on corpora similarity methods. Some measures have been proposed –mostly based on word frequency measures–, but they have not been sufficiently evaluated and indeed there is no standard measure. And regarding corpora in different languages, there is no precedent for measuring similarity.
3. One of the things to be tried is to see whether manual revision of the translated vectors to be used in the domain filtering yields a better performance.
4. Furthermore, for monitoring the improvements in the methodology, we intend to make tests with more corpora and to perform further research on multilingual corpora similarity methods.

REFERENCES

[1] Marco Baroni and Silvia Bernardini. 2004. BootCaT:Bootstrapping corpora and terms from the web.Proceedings of LREC 2004, 1313-1316. ELRA, Lisbon, Portugal.

[2] Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. Proceedings of HLT/NAACL 2003, 16-23. NAACL, Edmonton, USA.

[3] Božo Bekavac, Petya Osenova, Kiril Simov and Marko Tadić. 2004. Making Monolingual Corpora Comparable: a Case Study of Bulgarian & Croatian. Proceedings of LREC 2004, 1187- 1190. ELRA, Lisbon, Portugal.

[4] Martin Braschler and Peter Schäuble. 1998. Multilingual information retrieval based on document alignment techniques. Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, 183-197. Springer, Heraklion, Greece.

[5] Andrei Z. Broder. 2000. Identifying and filtering near-duplicate documents. Proceedings of Combinatorial Pattern Matching: 11th Annual Symposium, 1-10. Springer, Montreal, Canada.

[6] Adam Kilgarriff. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. Proceedings of workshop on very large corpora, 231-245. ACL SIGDAT,

[7] Beijing and Hong Kong, China. Michael D. Lee, Brandon Pincombe and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. Proceedings of CogSci2005, 1254-1259. Earlbaum, Stresa, Italy.

[8] Igor Leturia, Iñaki San Vicente, Xabier Saralegi, Maddalen Lopez de Lacalle. 2008. Basque specialized corpora from the web: languagespecific performance tweaks and improving topic precision. Proceedings of the 4th Web as Corpus Workshop, 40-46. ACL SIGWAC, Marrakech, Morocco.

[9] Igor Leturia, Antton Gurrutxaga, Nerea areta, Eli Pociello. 2008. Analysis and performance of morphological query expansion and languagefiltering words on Basque web searching. Proceedings of LREC 2008. ELRA, Marrakech, Morocco.

[10] Igor Leturia, Antton Gurrutxaga, Iñaki Alegria and Aitzol Ezeiza. 2007. CorpEus, a ‘web as corpus’ tool designed for the agglutinative nature of Basque. Proceedings of the 3rd Web as Corpus workshop, 69-81. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

[11] Igor Leturia, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria and Aitzol Ezeiza. 2007. EusBila, a search service designed for the agglutinative nature of Basque. Proceedings of Improving non-English web searching (iNEWS’07) workshop, 47-54. SIGIR, Amsterdam, The Netherlands.

[12] Emmanuel Morin, Béatrice Daille, Koichi Takeuchi and Kyo Kageura. 2007. Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 664-671. ACL, Prague, Czech Republic.

[13] Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics, 31(4):477-504.

[14] Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. Proceedings of the 37th annual meeting of the Association for Computational Linguistics, 519-526. ACL, College Park, Maryland, USA.

[15 Igor Leturia, Iñaki San Vicente, Xabier Saralegi Elhuyar Fundazioa R&D Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet Zelai Haundi kalea, 3. Osinalde Industrialdea 20170 Usurbil. Basque Country

[16] Luís Teixeira, Gabriel Lopes, and Rita A. Ribeiro Automatic Extraction of Document Topics CA3-Uninova, FCT, Universidade Nova de Lisboa 2829-516 Caparica, Portugal

[17] Adam Kilgarriff Comparing Corpora : International Journal of Corpus Linguistics, 6:1 (2001), 97–133.

[18] Young Mee Chung* and Jae Yun Lee , A Corpus-Based Approach to Comparative Evaluation of Statistical Term Association Measures

[19] Pimwadee Chaovalit, Lina Zhou, Movie Review Mining: a Comparison between Supervised and Unsupervised Classification ApproachesProceedings of the 38th Hawaii International Conference on System Sciences - 2005

[20] G.Vinodhini, RM.Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, International Journal of Advanced Research in Computer Science and Software Engineering,

Authors Biography

P.Santhi Priya received B.Tech(Computer Science and Engineering) from Jawaharlal Nehru Technological University,Hyderabad in 2003, M.Tech from JNTUH,Hyderabad in the year 2009. She is doing a part-time research in University College of Engineering & Technology,Acharya Nagarjuna University,Guntur,A.P. She worked as Associate Professor in the Department of CSE in GRIET, Hyderabad

Dr.T.v.Rao, B.E-ECE, M.E-CS, Ph.D in computer science and engineering, Wayne State University, Detroit, USA, Currently working as Professor in KL University, Vaddeswaram,Guntur Dt. He has more than 32 years of experience and has published many papers in national and international conferences. His areas of interest are multicore and parallel programming