2017 International Conference on Applied Mechanics and Mechanical Automation (AMMA 2017)
ISBN: 978-1-60595-471-4

Concept Extraction Based on Hybrid Approach Combined with Semantic Analysis

Xian-ming YAO[1], Jian-hou GAN[2] and Jian XU[1,*]

[1]School of Information Engineering, QuJing Normal University, Qujing, Yunnan, China

[2]Key Laboratory of Educational Informatization for Nationalities (YNNU), Kunming, Yunnan, China

*Corresponding author

Keywords: Concept extraction, Domain relevance, Domain consensus, Semantic analysis.

Abstract. This paper proposed a hybrid approach for concept extraction, which combined linguistic, statistical and semantic analyses. Similar to traditional methods, linguistic and statistical analyses were reserved; this paper innovatively adopted semantic analysis to analyze the semantic relationships of concepts in order to get a fine grained result. Another creative work is the combination of several statistical approaches from different perspective to filter the candidate concepts. Experiments had shown a promising result.

Introduction

The importance of concept extraction had been widely introduced in other papers [1,2]. As a basic task for ontology construction, it had been studied for a long period, still promising result could not achieve. This is mainly due to the ambiguous problems, such as the definition of concept, the evaluation, etc. So the improvement of precision remains a challenging task.

Nowadays, numerous techniques and tools had been proposed to overcome these problems. Lexicon and dictionaries were firstly borrowed as domain terms directly [3,4,5], the lexical and syntactic patterns were exploited to the construction of ontology, but the lexicon resources are rare and mainly constructed for general purpose. Rule based [6,7], NLP based [8,9,10,11,12,13], and statistical [14,15] based techniques were put forward to automatically or semi-automatically extract concepts. The desire to extract concepts from free text [3] was arising with the huge electronic text available from internet. Rule based techniques, such as upper cases of letter, "such as ……", and "is a" could be borrowed to extract term and relationship [12]. It could reach high precision while recall is lower. NLP techniques utilize lexical and syntactic analysis [13] to determine the role terms plays in sentence, and patterns of POS could be designed to affiliate extraction [11]. Statistical techniques, such as TF, TF-IDF, etc. [16,17], had been studied for a long history, and applied to term extraction around 1980s, and got a better result. TF-IDF, as the widely used technique, was introduced to measure the importance of terms to a domain [18]. In order to achieve more accurate result, hybrid approaches which combining NLP techniques and statistical method were adopted [4,17,19]. NLP techniques were used to select candidate concepts satisfying syntactic structure, and statistical techniques were used to measure the degree which candidate concept belongs to a specific domain.

Although, works done before had got a good result, two important problems had been neglected:

(1). the nature of concept had not been taken into consideration comprehensively. Statistical approaches, such as TF-IDF, C-Value [18,20], etc. had been proposed to compute Domain Relevance or Lexical Cohesion [19,21,22] separately, but comprehensive considerations of all the nature of concept were merely mentioned. This paper introduced two kinds of operational nature: Domain Relevance, Domain Consensus. Different statistical method was adopted to measure its strength.

(2). Semantic relationships between concepts were neglected [19,23]. NLP and Statistical method mainly focus on the characteristic of term itself, the semantic relations which meaningful to ontology had not been taken into consideration. We believe that it plays an important role to concept extraction. Word co-occurrence was adopted to measure the semantic relationship and give proposal to new concept selection.

Concept Extraction Framework

A premise should be put forward firstly: as a linguistic realization, concept extraction is often being simplified as term extraction in a not strict manner [18,22,24,25,26] . This paper mainly focus on terms, neglected the influence of synonym, antonym, abbreviation etc. The remaining work will be studied in future.

The concept extraction framework is depicted in Fig.1 as bellow. Rectangles represent data flow from left to right. Domain corpora were collected from various ways. In the first phase, Candidate concepts could be extracted by using syntactical patterns. In the second phase, statistic filtering methods could be used to measure some kinds of degree of candidate concepts. In the last phase, semantic analysis was used to determine the semantic relationships between concepts, and related candidate concepts were collected as the domain concept. Rounded rectangles in the figure represent techniques used in different phases.

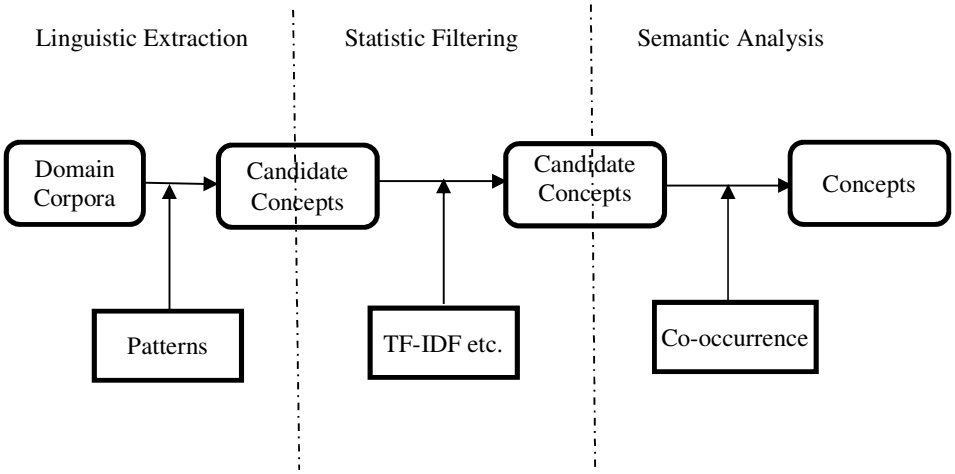

Figure 1. Concept extraction framework.

From the foregoing, we could know there are three phases composed concept extraction framework: Linguistic Extraction, Statistic Filtering, and Semantic Analysis. For each phase, specific tasks were given bellow.

Linguistic Extraction

Linguistic Extraction accomplishes the task of extracting candidate concepts from free text. Generally speaking, concepts usually composed by nouns. We could utilize the law to extract candidate concepts. Manually designed linguistic patterns showed in Table 1 were used to match the POS. The closed pattern could promote precision [20].

Table 1. Syntactic patterns.

Noun	n, ng, vn, vg
Noun +Noun	RandComb(Noun)
Modifier +Noun	RandComb(a, ag+v, b)+RandComb(Noun)

In order to accomplish the task, we borrowed ICTCLAS2016 tool to segment sentences into words and tag part of speech (POS). Each word is assigned a POS, such as "n", "vn". We could extract those words with POS mentioned in Table 1. Single Nouns could be extracted with POS of "n" or "ng", or "vn", or "vg" as terms. Compound Nouns could be extracted as terms with random combination of any length and any number of the four kinds of POS. Compound Nouns with modifier, such as "a", "ag+v", "b", could also be extracted as terms [11, 12].

When terms extracted out, sequential index file (SIF) and inverted index file (IIF) with distribution information will be created to realize fast access of terms.

Statistic Filtering

Statistic Filtering accomplishes the filtering of candidate concepts according to four kinds of features. Paper [20] had reported that combined statistic method could improve precision rate with little loss of recall rate.

Linguistic approaches were never used alone. The candidate concepts contain too much error, and statistic approaches are needed to filter those irrelevant. The commonly used metrics were Domain Relevance and Domain Consensus. We would like to introduce Domain Relevance, Domain Consensus here.

(1). Domain Relevance measures the widely use of term in domain corpora, but merely used in other domains. This is a clue of active interaction of terms to domain words, and is a preferable measure to domain concept. It could be defined as: For term t, Domain Relevance to a specific domain D_i in Domains (D_1, D_2,……, D_n) is:

$$\mathrm{DR}(\mathrm{t}, \mathrm{D}_i) = \frac{P\left(\frac{t}{\mathrm{D}_i}\right)}{\underset{1 \le j \le n}{maxP}\left(\frac{t}{\mathrm{D}_j}\right)} \tag{1}$$

(2).Domain Consensus measures the distributed use of a term in a domain [21]. If term consistently used across domain documents with high frequency [27], it could be regarded as domain concept. Domain Consensus could be defined as: For term t in domain documents D=(d_1,d_2,……, d_n):

$$\mathrm{DC}(\mathrm{t}, \mathrm{D}) = \sum_{\mathrm{d}_j \in D} P(t, \mathrm{d}_j) log_2^{\left\{\frac{1}{P(t, \mathrm{d}_j)}\right\}} \tag{2}$$

Except for the natures mentioned above, there are still other nature, including Structural Relevance, and Miscellaneous [27], but unfortunately, they are not operational, so they had not been taken into consideration. Another important nature is Lexical Cohesion[27,28,29] which measures terms with length above 2.In this paper, we hasn't adopted, for concepts extracted out shows that terms are almost short.

To coordinate the factors, a simple equation is given as bellow, and coefficients were assigned to each factor which denotes its importance. The final score to each candidate term is defined as:

$$\mathrm{Score}(\mathrm{t}) = \alpha \mathrm{DR}(\mathrm{t}, \mathrm{D}_i) + \beta \mathrm{DC}(\mathrm{t}, \mathrm{D}) \tag{3}$$

Initially, α, β were set as 0.5 [21]. It should be refined according to precision and recall empirically.

When Score for each term was calculated, a descendant list ranked by score would be given out, and a threshold would be used to cut those irrelevant terms.

Semantic Filtering

Hybrid approaches combine linguistic and statistical methods mentioned above dominated the mainstream in the past and relatively higher precision could reached according to literatures. But it mainly focused on terms itself, failed to make a comprehensive study for semantic relatedness between terms, which is important for organic ontology. This paper tried to implement semantic analysis upon traditional hybrid approach.

Although plain documents were sequential, terms hidden behind is highly correlated with each other. Semantically related terms seem to appear in the same context, or topics, even sentences. Distribution Hypothesis could be used to explain the phenomenon. It is manifested at word co-occurrence [29,30,31] and collocation. Word co-occurrence is widely used in information retrieval, word disambiguation, etc. and also term extraction. Here we did not use word co-occurrence as a statistic filter, but a semantic analysis tool to select terms with highly semantic relationship.

Our Semantic Filtering framework is illustrated as Fig.2 shown. Similar to Fig.1, rounded rectangles indicate data flow, and rectangles represent techniques adopted. Candidate concepts were extracted from the previous two phases. The final product is domain concepts.

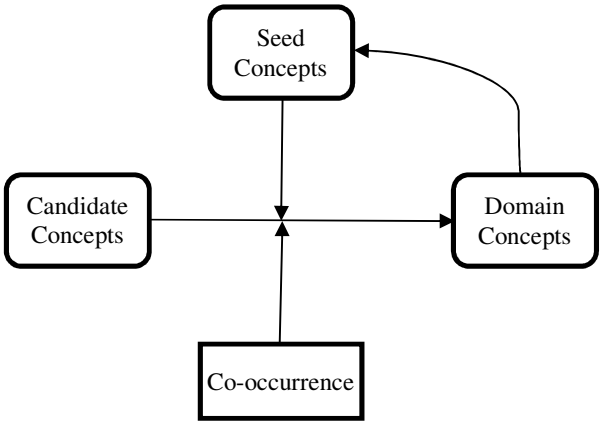

Figure 2. Semantic filtering framework.

The filtering loops could be described as bellow.

Firstly, seed concepts and candidate concepts were prepared. Seed concepts were manually selected and provided, while candidate concepts were got from previous phase.

Secondly, word co-occurrence was adopted to find search candidate concepts which appear in the same documents with seed concepts.

Thirdly, the relatedness between seed concept and candidate concepts should be calculated.

Let T_s be the seed concept, and $T_c=(t_1, t_2, \ldots\ldots, t_n)$ be the candidate concepts. They have two kinds of frequency:

Absolute Frequency which is overall occurrence of concept related to domain corpora, defined as $T_{s_i}^{a}$ and t_i^a separately.

Relative Frequency which is co-occurrence related to seed term, defined as t_i^r.

To a specific candidate term t_i, its strength of semantic relatedness with T_s could be described with Jaccard Coefficient and defined as:

$$S_i = \frac{t_i^r}{T_{s_i}^{a}+t_i^a-t_i^r} \tag{4}$$

S_i could be considered as semantic strength of relatedness, and it is normalized to the range of (0, 1). Candidate concepts T_c were ranked descendant according to S_i, and also an empirical threshold was given to filter candidate concepts that with lower semantic relationship.

Thirdly, new concepts computed from above were added into seed concept library, and do loop to the first step until no new domain concepts could be found.

The result of semantic filtering is a set of domain terms that closely related to each other. It could be used as domain concept directly to support the construction of domain ontology.

Evaluation

In order to validate the effectiveness, we implemented the algorithms to the domain of cultural heritage preservation for Yunnan China minorities.

Yunnan Minority Ontology Construction Toolkit

An automatic ontology construction toolkit, call Yunnan Minority Ontology (YMO), were developed at the same time. It is aimed at providing a unified and simple way for automatic domain ontology construction. A simply designed website crawler was developed and gathered 4247 relevant files with a size 7.04M. In addition, ICTCLAS2016 were borrowed to accomplish POS tag tasks, and textforever which is a text process software was used to do some text preprocessing work.

Linguistic Selection

To accomplish this task, syntactical patterns were designed to select candidate terms from plain text. Here, we mainly adopted POS information. After analysis to text, we choose words with POS of "n",

"ng", "vn", "vg", and its random combination to phrases. Nouns with concrete meaning, such as location, person names, etc. were not in consideration for they usually stand for an instance of concept. Three kinds of modifiers and their random combination to candidate terms were adopted as rules, these POSs are "a", "ag+v", "b".

These rules work effectively with some wrong terms in the domain of Minority. When they were applied to other domains at the same time, unexpectively, they are especially effective in the domain of Law. This is mainly due to that the Law is strictly designed, while texts in the domain of Minority are arbitrary.

After comprehensive analysis, we find that, lexical patterns could be adopted to improve precision and recall rate. Some wrapper indicators, such as "<< >>",".", etc. have special meanings.

The greatest challenge lays that domain terms usually consisted in the form of "v+n", such as: "*Chui1chui1qiang1*", "*Da3ge1jie2*", "*Cai3gu3jie2*". (Note that words started with an upper letter and composed of characters and numbers with italic font in this paper represents a Chinese word. English characters represent its spelling, while number represents the tone before it.) According to the rules, these terms could not be extracted, for verbs had not been adopted. Once adopted, huge number of wrong term would be extracted.

Statistic Filtering

The candidate terms extracted above were sending into the second phase to calculate Domain Relevance (DR) and Domain Consensus (DC). The length of term was taken into consideration as well, it is especially important to long terms. The computation result is shown in Table 2.

Table 2. Statistics of terms.

Term	AbsFreq	DR	Docs	DC	TermLen	Final Score
Ren2	9087	0.40	2761	0.99	0.1	0.54
Shi2	3172	0.61	1642	0.86	0.1	0.49
Min2zu2	2758	0.92	1251	0.67	0.2	0.43
Di4	1263	0.77	847	0.58	0.1	0.37
Di4qu1	2383	0.60	1125	0.60	0.2	0.36
Li4shi3	1544	0.53	929	0.61	0.2	0.36
Ren2men1	2070	0.53	1060	0.60	0.2	0.35
Ren2min2	1661	0.77	905	0.51	0.2	0.33
Di4fang1	1149	0.46	750	0.52	0.2	0.30
Fu4nv3	1906	0.93	799	0.41	0.2	0.30
Cun1zhai4	1145	0.96	669	0.40	0.2	0.29
Wen2hua4	1820	0.79	721	0.41	0.2	0.29
Shui3	1476	0.58	738	0.45	0.1	0.29
Jie2ri4	1405	0.96	711	0.37	0.2	0.28

The first column represents candidate terms.

The second column represents absolute frequency of each term in domain corpora, this value could be used to calculate Domain Relevance, and also it could provide an intuitive impression to us.

The third column represents the domain relevance of each term. In order to calculate DR, we collected three kinds of other domain corpora: The Law, The Culture, and The Science. This value is calculated according to Eq.1

The fourth column represents the number of documents that terms occurred. This value will be used to calculate DC, and also an intuitive value.

The fifth column represents the value of DC. The value was calculated according to Eq.2 and normalized to the range of (0, 1).

The sixth column represents the value of term length and it is multiplied by a double value 0.1. This value was introduced in afterward. We find that terms with long length which have definite meaning and important to domain can hardly be extracted, so we assign a value according to its length.

The last column represents the final score for each candidate terms. The score was calculated according to Eq.3. The values of α, β, γ were assigned as 0.1, 0.5, 0.4 separately. The ultimate equation of Eq.3 is as follows:

$$\text{Score(t)} = 0.1 * \text{DR} + 0.5 * \text{DC} + 0.4 * \text{TermLen} \quad (5)$$

Reasons that α=0.1 is: according to our analysis to DR, we found that numerous terms that occur only once in domain corpora gain the highest score of DR, the score reached as high as 1.0. For they never occur in other domains. This would lead that the final score will be mainly depending on domain relevance.

Reasons that β=0.4 is: most terms is as low as 0.1 while the have a good Domain Consensus representation.

Finally, the value of DR and DC were normalized to a range of (0, 1).

Semantic Analysis

Semantic analysis is especially important to concept extraction, for concepts in domain corpora are closely related to each other. As mentioned above, we adopt word co-occurrence to accomplish this task. Word co-occurrence require researcher provide some seed concepts, and it will choose one of them to search new candidate concepts, and candidate concepts that satisfy an threshold will be added to concept library.

After analysis to term frequency, domain relevance etc. we selected the following terms as seed concepts: (Shao3shu4min2zu2, Jie2ri4, Cun1zhai4, Wen2hua4, Chuan2tong3, Yi2shi4, Wu3dao3, Min2zu2te4se4, Chuan2tong3jie2ri4, Zong1jiao3xin3yang2, Feng1su2xi2guan4, Jin1ji4, Yin1yue4). These seed concepts possess a good domain representation; they exist merely in other domain, and spread widely across domain corpora. In the meanwhile, they are a good indicator to different aspects which we are interested in.

With these seed concepts, we could extract other candidate concepts which co-occur with them; here, we restricted candidate concepts that co-occur in a level of documents, but not paragraphs or sentences [31]. Paper [29] had reported the effectiveness in this level; also, each document we collected is mainly focus on a topic, such as festival, music, tradition, etc.

Equation 3.4 is borrowed to evaluate the relatedness of candidate concepts and seed concept. One of the calculated results is listed in Table 3.

Table 3. Mutual information.

Seed Concept	New Concepts						
Yin1yue4	*Min2ge1*	*Ge1*	*Qu3diao4*	*Gu3yue4*	*Wu3dao3*	*Jie2zhou4*	*Yue4qi4*
MI	0.2775	0.2738	0.2391	0.2310	0.2074	0.1973	0.1921

It is noteworthy that Mutual information (MI) is not in the range of (0,1), when seed concepts with lower term frequency and occur in fewer documents, those candidate concepts which spread widely across domain corpora could gain a score larger than 1. But, the larger score it gains, the least relatedness with seed concept, it could be a general concept. A strategy is reverse and normalize MI to the range of (0, 1), and merge them with those candidate concepts with MI lower than 1.0, this process is just like fold the page of your book. Consequently, all the MI value is reasonably normalized to the range of (0, 1).

Candidate concepts with MI value closest to 1.0 are more likely to be domain concept, and an empirical threshold was given to restrict the number of new concepts. We set the value as 0.15, only candidate concepts with MI higher than 0.15 would be selected. Some of interesting extraction result is shown in Table 4.

From the table, we could get an exciting extraction result: concepts are closely related to each other. Although, instances in the table are limited, analysis to the whole extraction process shows that the result is promising.

Table 4. Extraction result.

Seed Concept	New Concepts
Shao3shu4min2zu1	*Min1zu2, Ren1kou3, Wen1hua2, Xiang1,Yu3yan2, Guo1jia1, Li4shi1, Fu1shi4*
Ge1wu3	*Wu3dao3, Ge1, Min1ge1, Lu2sheng1, Jie2zou4, Guo1*
Chuan2tong3jie2ri4	*Jie2, Jie2ri4, Nong2li4, Zang4li4, Huo2dong4, Huo3ba3jie2, Lu2sheng1*
Yu3yan2	*Wen2zi4, Yu3, Min2zu2, Wen2, Fang1yan2, Ge1, Shui3zu2, Wen2hua2, Ren2kou2*
Xin1niang2	*Xin1lang2, Nv3fang1, Nan2fang1, Hun1li3, Gu1niang1, Mei2ren2*
Ke4ren2	*Jiu3, Zhu3ren2, Xin1niang2, Cha2, Rou4, He1*
Guan1cai2	*Si3zhe3, Guan1, Shi1ti3, Fen1, Chan3fu4*
La2ma2	*Ha3da2, Huo2fo2, Shi1ti3, Jin4ji4, Gu3hui1*
Si4yuan2	*Shen2shan1, Si4, Fo2xiang4, Jiao4, Qing1zhen1si4,Fo2si4, Xin4tu2*

Evaluation

In order to assess the overall extraction effect, we had made a comprehensive comparison between different phases. Precision rate, recall rate, and F-measure [32] were borrowed. Table 5 shows the result.

Table 5. Overall performance.

	Pattern	Statistic	Semantic
Precision rate	40.9%	75.3%	81.4%
Recall rate	90.7%	78.5%	74.0%
F-measure	56.4%	76.9%	77.5%

From the table above, we could notice that precision rate is getting better, while the recall rate is getting bad, it suggests that pattern based term extraction is a bottleneck. Good news is that the overall F-measure is improving. The result shows the effectiveness of our algorithm.

Summary

In this paper, we adopted a hybrid approach which combines pattern, statistic, and semantic methods. This is approach is a little different from traditional approaches which combines pattern and statistical methods. A semantic analysis was adopted to collect concepts closely related to each other and a relatively higher precision could be got.

When we go deeper, several problems emerged:

1. Rules had become the bottleneck to term extraction, new strategies should be put forward to deal with arbitrary text;
2. Whether machine learning techniques could improve candidate terms extraction effect? [2,34,35]
3. The impact of named entities could not be determined, they always co-occur with concepts and is helpful to concept extraction;
4. Although several strategies had been applied, extraction for phrases seems to be failed;
5. New concept selection strategies in word co-occurrence should be improved, the strategy we adopted tends to select those with lower DC and DR value which do not have a good domain representation.

The problems we mentioned above are helpful to concept extraction, and we will try to solve in future work.

Acknowledgement

The research is supported by the National Nature Science Fund Project (61562093), Key Project of Applied Basic Research Program of Yunnan Province(2016FA024).

References

[1] Zouaq, Amal, and R. Nkambou. A Survey of Domain Ontology Engineering: Methods and Tools. Advances in Intelligent Tutoring Systems. Springer Berlin Heidelberg, 2010:103-119.

[2] Wong W, Liu W, Bennamoun M. Ontology learning from text: A look back and into the future, J. Acm Computing Surveys, 44(2012)1-36.

[3] Dahab, Yehia M, Hassan, et al. TextOntoEx: Automatic ontology construction from natural English text, J. Expert Systems with Applications, 34(2008)1474-1480.

[4] Navigli R, Velardi P, Cucchiarelli A, et al. Quantitative and qualitative evaluation of the OntoLearn ontology learning system, C. International Conference on Computational Linguistics. Association for Computational Linguistics, 2004:1043.

[5] Velardi, P., Navigli, R., Cucchiarelli, A., and Neri, F. 2005. Evaluation of OntoLearn, a methodology for automatic learning of ontologies. In Ontology Learning from Text: Methods, Evaluation and Applications, P. Buitelaar, P. Cimmiano, and B. Magnini, Eds. IOS Press, Hershay, PA.

[6] Vassileva J, Deters R. Dynamic Courseware Generation on the WWW, J. British Journal of Educational Technology, 29(1998)5-14.

[7] Stojanovic L, Stojanovic N, Volz R. Migrating data-intensive web sites into the Semantic Webm, C. Proceedings of the 2002 ACM symposium on Applied computing. ACM, 2002:1100-1107.

[8] Wermter J, Hahn U. Finding new terminology in very large corpora, C. International Conference on Knowledge Capture. 2005:137-144.

[9] Sclano F, Velardi P. TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. Enterprise Interoperability II. Springer London, 2007, pp. 287-290.

[10]M Sabou, Learning Web Service Ontologies: an Automatic Extraction Method and its Evaluation, Ontology Learning from Text Methods Evaluation & Application , 2005.

[11]Dan I. Moldovan, Roxana C. Gîrju. An Interactive Tool for the Rapid Development of Knowledge Bases, J. International Journal on Artificial Intelligence Tools, 10(2001)65-86.

[12]ML Reinberger, P Spyns, Unsupervised Text Mining for the learning of DOGMA-inspired Ontologies, Ontology Learning from Text Methods, 2005.

[13]Zouaq A, Nkambou R. Enhancing Learning Objects with an Ontology-Based Memory, J. IEEE Transactions on Knowledge & Data Engineering, 21(2009)881-893.

[14]Turney P D. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, C. European Conference on Machine Learning. Springer-Verlag, 2001:491-502.

[15]Alexander Budanitsky. 1999. Lexical Semantic Relatedness and its Application in Natural Language Processing, technical report CSRG-390, Department of Computer Science, University of Toronto, August 1999. http://www.cs.toronto.edu/compling/Publications/ Abstracts/Theses/ Budanitsky-thabs.html

[16]Salton G, Buckley C. Term-weighting approaches in automatic text retrieval, J. Information Processing & Management, 24(1988)513-523.

[17]Pazienza M T, Pennacchiotti M, Zanzotto F M. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches, J. Studies in Fuzziness & Soft Computing, 185(2005)255-279.

[18]K Frantzi , S Ananiadou , H Mima, K. Frantzi, S. Ananiadou, H. Mima, Natural language processing for digital libraries Automatic recognition of multi-word terms: the C-value/NC-value method,Int. J. Digit. Libr. 3 (2000) 115–130.

[19]Navigli, Roberto, Velardi, et al. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites, J. Computational Linguistics, 30(2004)151-179.

[20]Frantzi K T, Ananiadou S, Tsujii J. The C-value/NC-value Method of Automatic Recognition for Multi-word Terms, C. European Conference on Research and Advanced Technology for Digital Libraries. Springer-Verlag, 1998:585-604.

[21]Velardi P, Missikoff M, Basili R. Identification of relevant terms to support the construction of domain ontologies, C. The Workshop on Human Language Technology and Knowledge Management. Association for Computational Linguistics, 2001:5.

[22]Pazienza M T. A domain-specific terminology-extraction system, J. Terminology, 5(1998) 183-201.

[23]Haggag M H. Keyword Extraction using Semantic Analysis, J. International Journal of Computer Applications, 61(2013)1-6.

[24]Sager J C, Dungworth D, Mcdonald P F M A. English special languages : principles and practice in science and technology[M]. Brandstetter, 1980.

[25]P. Buitelaar, P. Cimiano, B. Magnini, Ontology learning from text: an overview, in: P. Buitelaar, P. Cimiano, B. Magnini (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands, 2005, pp. 1–10.

[26]Jacquemin, C. Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. PhD thesis, Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes (1997)

[27]Fortuna B, Lavrač N, Velardi P. Advancing Topic Ontology Learning through Term Extraction, C. Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence. Springer-Verlag, 2008:626-635.

[28]Morris J, Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text, J. Computational Linguistics, 17(1991)21-48.

[29]Ohsawa Y, Benson N E, Yachida M. KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor, C. Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on. IEEE, 1998:12-18.

[30]Wartena C, Brussee R, Slakhorst W. Keyword Extraction Using Word Co-occurrence, C. Database and Expert Systems Applications. IEEE, 2010:54-58.

[31]Yutaka Matsuo, Mitsuru Ishizuka. Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information, J. International Journal on Artificial Intelligence Tools, 13(2003)157-169.

[32]Salton G, Mcgill M J. Introduction to modern information retrieval. McGraw-Hill, 1983.

[33]Wong W, Liu W, Bennamoun M. Ontology learning from text: A look back and into the future, J. Acm Computing Surveys, 44(2012)1-36.

[34]Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources, J. Journal of the American Medical Informatics Association Jamia, 18(2011):580.

[35]Melli G, Ester M. Supervised identification and linking of concept mentions to a domain-specific ontology, C. ACM International Conference on Information and Knowledge Management. ACM, 2010:1717-1720.