International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.11 (2015), pp.423-432
http://dx.doi.org/10.14257/ijsip.2015.8.11.38

# A Multi-Domain Web Text Feature Extraction Model for e-Science Environment

Weng Yu

*School of Information Engineering, Minzu University of China, Beijing, China*
*dr_wengyu@126.com*

***Abstract***

*The traditional information extraction methods based on specific domain usually depend on the domain dictionaries to discover the text feature. It is inconvenient for reproducing and difficult to transplant in multi-domain environment. The application scope is limited seriously. Oriented to the deficiencies above, a multi-domain web text feature extraction model for e-Science is proposed (named e-FTM). This model adopts the Chinese split words technology without dictionary into the process of multi-domain text feature discovery and avoids the dependency of domain dictionaries effectively. With the help of classification of common and individual features, the model tracks the generation and the development trend of domain events dynamically, and forms a couple of local data centers eventually. Through cooperative scheduling the domain knowledge between different local data centers, the knowledge utilization efficiency of the domain information in the global scope is improved sharply. To validate the performance, the experiments on the multi-domain text feature extraction, topic features dynamical tracking and the domain knowledge cooperative scheduling demonstrate that the model has higher application validity and practicality in e-Science environment.*

***Keywords:*** *e-Science Environment; Feature Discovery; Multi-Domain Data Model; Web Text Mining*

## 1. Introduction

The multi-Domain web text feature extraction model for e-Science environment aims to use the technologies of text mining to categorize the massive web data, to organize and use the scientific information in web data effectively and reasonably, to establish e-Science service platform of field science data. The platform will effectively improve the portability, feature classification and domain knowledge's cooperative ability and sharing ability of the information extraction model in the multi-Domain environment provide data support for different domain experts' varieties of complex applications.

As a typical data-intensive application, web text feature extraction model for e-Science environment has the following characteristics:

The user can manipulate and use the data from any point on the Internet, though they are dispersed geographically.

The modal collected the web data from the portal website, research sites, BBS and other Information interaction platform in all domains incrementally by means of the web crawler technology, the storage and computing of data relatively concentrated, and as time goes on, it gradually becomes the data center of many domains.

On-demand domain. Users from different domains can find and extract the scientific data which have domain features based on personal application requirements.

A large number of domain characteristic words or phrases make the model difficult to find and extract the web data features in the multi-Domain environment by traditional extraction method based on information of the dictionary [1].

ISSN: 2005-4254 IJSIP

Users from the same domains who have the similar feature requirements hope to collaborative share the domain knowledge found by other researchers, so the domain information turn from localized into global.

Focusing on above application features, this article has presented multi-Domain web text feature extraction model for e-Science environment (named as e-FTM).This model could find and extract the web data features in multi-Domain environment, in the meantime, it collaborates and shares the knowledge from the same domain, tracks the change of every domain features dynamically, forms the information database of domain features gradually. To provide a powerful guarantee for the complex data in multi-Domain environment, the main contributions as follows:

1．The web features extraction algorithm in multi-Domain environment is put forward. The traditional text features extraction method mostly relied on domain dictionary to realize the word segmentation of the target text, this method is simple and feasible for the specific domain knowledge, however it is hard to fit in with the text features extraction in multi-Domain environment. With the progress of domain knowledge, some new words are hard to find in time with the traditional method, e-FTM synthesizes the above shortage, puts the word segmentation technology with no dictionary into the web text feature discovery, abandons the dependence of domain dictionary and improves the portability.

2．A series of web text features dynamic tracing method is established. As a web text features extraction model, e-FTM regularly monitors the updates of the Web topic and the features of his event in multi-domain environment, mining the features changing and development trend of different topic events, establish the knowledge database of dynamic domain features, provide data support for a variety of complex data-intensive applications.

3．The collaboration and sharing policy of a series of domain knowledge is Put forward. In order to improve the knowledge utilization of the model under the e - Science environment, e-FTM abstractly describes the domain events features that found by local monitoring nodes, establish regional domain features cases (such as high frequency event features cases, new event features cases and so on). E-FTM collaborative scheduling between different monitoring nodes of domain features cases, effectively promote the discovery and sharing ability of the domain knowledge in global scope.

This article mainly consists of the following sections: Part 2 introduces the related research work on web feature extraction; part 3 generally describes the issues involved and working process of model; part 4 describes the calculating process of the web text feature extraction model for e-Science environment in details; Part 5 verifies the validity of the model; part 6 draws the final conclusion and the future research work.

## 2. Related Works

In recent years, with the rapid development of information technology, web text mining techniques have attracted more and more attention of experts and applications. Khaled Khelif [1] (2007) proposed a method for ontology-based information extraction, intended to help biologists to get professional knowledge more effectively. This method relies on the semantic annotation of scientific literature, generate the domain ontology automatically and provide the retrieval interface of appropriate information. Tara McIntosh [2] (2007) proposed a full text information extraction model oriented to biomedical based on insufficient abstracts of traditional analytical methods. Ziya Ozkan Gokturk and Nihan Kesim Cicekli *et al.*, [3] (2007) resorted the web crawler technology, used preset regex to achieve the extraction and classification of web pages metadata. Experiment with the European Cup and UEFA Champions League as examples, regularly download specified sport site information, extracted event information metadata, and then taped the latest sporting events. Rodney D. Nielsen and Wayne Ward *et al.*, [4] (2008) combined with the actual needs of the automatic teaching, proposed a semantic representation of the text and proved its effectiveness. Veronica Dahl and Baohua Gu [5] (2006) described a text processing method for biomedical analysis and its associated concepts. This approach took semantic characteristics of different medical concepts and grammatical constraints in expression as classification criteria, and realized the feature extraction of phrases while improving fault-tolerance ability for irregular text. B. Martins and H.Manguinhos *et al.*, [6] (2009) aiming at the metadata representation abnormal phenomenon in geographic information systems (for example: data incomplete or machine unreadable), introduced a time-based representation information extraction model. The model with Web gazetteer, using a relatively simple way to extract geographic time information, combined with space time, to get a more comprehensive description of geographic information metadata. Honglei Guo and Huijia Zhu [7] (2009) present a named entity detection model based on semantic association analysis. The model effectively overcame the differences of the distribution of data among different domains by mining the latent semantic association between words, contributed to the accuracy of entity identification. Rajib Verma [8] (2009) put the web text mining technology into Internet users' sentiment analysis, and provided data support for a variety of complex applications for financial, customer relationship, business and other domains. There are three representative analysis methods in Table 1:

**Table 1. The Comparison of Web Text Feature Extraction**

| Domain | Analysis | Performance |
|---|---|---|
| Medicine[9] | Extract the semi-structured data in Wikipedia through the semantic annotation methods, to build medical ontology automatically | The test set includes 4308 concepts and 7465 associated. With the continuous input of test set, the precision rate and recall rate can be maintained between the 70% -90%. |
| Social Network [10] | Analysis the sharing information between different entities of social networks, cluster the similar attributes to realize the latent semantic association extraction among entities | The test set includes 143 named entities in political network and 421 entities in research network, the precision rate is 76% and the recall rate is 68%. |
| Biomedical [11] | The feature selection method is applied to predict protein-protein interactions. | Hybrid Test Set：AIMed(2005), Bioinfer (2007), HPRD50 (2007) etc. The precision rate is 62% and the recall rate is 52%. |

These research methods mostly rely on the auxiliary of domain dictionary or marked word sets, although they can effectively improve the accuracy of specific domain feature extraction, however they cannot meet the actual demand in the multi-domain information extraction system portability. This article combines the application characteristics of the e-Science environment, puts the word segmentation technology without dictionary into the process of web text features discovery, gets rid of the dependence that traditional methods for domain dictionary effectively, and enhances the feature discover ability of the model in multi-domain environment. E-FTM dynamically tracks the status of each domain subject feature, and deeply mines the recent domain events and associated vocabulary, then forms the regional domain data centre gradually. It promotes collaboration and sharing by scheduling the latest domain knowledge in different monitoring node.

## 3. Problem Description and General Process

As a multi-domain text feature extraction model for e-Science environment, e-FTM mainly considers the following three aspects:

Firstly, the model gets rid of the dependence on the domain dictionary. The function of domain dictionary is text segmentation and data pretreatment for feature discovery in most of the Chinese information extraction models. But because of the limits of number and update speed, the capability of discovering new events and new domain vocabulary is seriously restricted, which is not conducive to the transplantation and extension for the model. The introduction of the segmentation technology with no dictionary will improve the knowledge study ability of the model effectively and make it more suitable for the feature's extraction and discovery of multi-domain text.

Secondly, e-FTM tracked the trend of development of the event characteristics. The events are not likely to immutable and frozen. They along with the development of domain knowledge, the event features tend to constantly update. At the same time of testing the event characteristics, e-FTM also give consideration to the track of the development trends , look back upon the origin of current characteristics, and provide reliable data support for a variety of application services (for example: the analysis of event correlation , the excavation of knowledge context).

Thirdly, e-FTM promoted the sharing of domain knowledge. The purpose of the e-Science is to be more effective to promote collaboration and sharing of scientific data between many experts and scholars [12]. Combined with the real demand of users, e-FTM will build a plurality of local area data center and dispatch the information of latest domain, through which to expand the detection speed and the utilization rate of knowledge of local emergency in the global range.

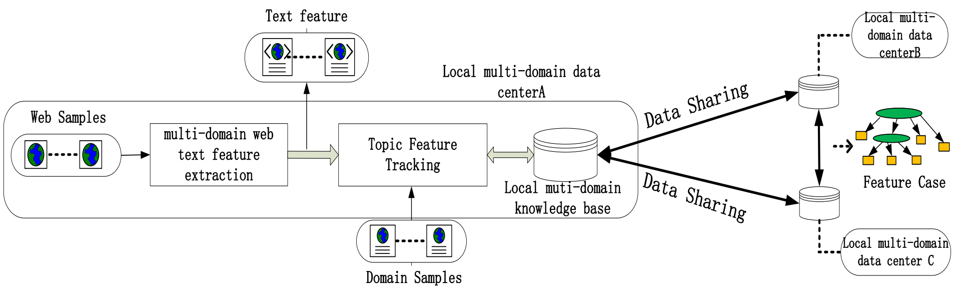


**Figure 1. The General Structure of e-FTM**

Overall architecture of e-FTM is shown in Figure 1. The model is composed by three parts, the Web text features discovery in multi-field dynamic tracking of topics feature, as well as collaboration and sharing of knowledge in the field. In the feature extraction process, domain experts provide the target texts according to the actual requirements. The model extracts the web texts which have the similar characteristics. By the updating of topic features constantly, e-FTM tracks the changing trends of topic features dynamically and gradually forms a regional multi-domain data centre. On the other hand, the model produces the historical cases from the characteristics of topics of local data centre, dispatches the latest features among the local data centres, and promotes discovery and sharing of the field knowledge in the global scope.

 

## 4. Multi-Domain Web Text Feature Extraction Model for e-Science Environment

### 4.1. Multi-Domain Web Text Feature Discovery

English use the blank as a segmentation words identifier which is different with Chinese. Chinese character is a form of expression derived from the hieroglyphics, each character has the independent meaning, and there is no clear separation between the words. Chinese needs the segmentation techniques to realize the automatic segmentation of the Chinese characters. For above reasons, most domain text feature extraction models rely on domain dictionary to realize the word segmentation of the target text. This method is simple and well applied to the discovery and extraction of the specific domain text features. However because of the limitations of their knowledge structure and update speed of the domain dictionary; it is not easy to promote and transplant into a multi-domain environment.

E-FTM put the word segmentation technology without dictionary into the discovery process of web text features to adapt to the needs of practical application environment.

Here, T is the given target text, domain experts provide objective sample set SampleDomain. (Note: the topic here only provide a sample target, rather than the specific domain dictionary, which will greatly increase the flexibility of the use of experts of the domain, reduce the complexity of application) FreqList and SplitWordsList are used to record the frequency of each character in the target text and the segmentation results. $\alpha_i$ and $\beta_j$ represent FreqList and SplitWordsList in the i-th element, where $\alpha_i = < Char, Freq >$, $\beta_j = < Word, Freq, Length >$.

TList is the final feature word string to output of the target text T. In order to further improve the accuracy of segmentation, the model sets up a disabled character called Stop CharacterList, used to store all punctuation and some no real intention of the auxiliary word. The specific details are shown in Table 2.

In step 1, the model statistics the frequency of the character to prepare for the subsequent word segmentation. However, for the disabled concentration by the emergence of the character, always maintain the frequency of 1.

In step 2, e-FTM put the character as the basic processing unit, extract the character one by one which is following the character and the frequency is 1, and record the length.

In step 3, e-FTM calculates the frequency of each feature string. The subset string frequency is calculated analyzed. Finally, sort the character according to the frequency and length of character strings. When two words have the same frequency, output the one whose length is longer preferentially.

**Table 2. Web Text Feature Extraction Algorithm for Multi-Domain**

| | |
|---|---|
| Input : | Target text T，domain target sample set SampleDomain, FreqList, SplitWordsList, StopCharacterList，$\alpha_i$ and $\beta_j$ |
| Output : | TList feature word list |
| Step 1: | Calculate the frequency of each character |
| | For each character $C$ of $T$ |
| | { $\alpha_i.char = C$ ; $\alpha_i.freq = 1$ ; |
| | Insert $\alpha_i$ into FreqList |
| | if( $\alpha_i$ StopCharacterList) { $\alpha_i.freq = 1$ } |
| | elseif(( $\omega$ Reverse（FreqList）)&&( $\omega$ == $\alpha_i.char$ )) |
| | { $\alpha_i.freq = \omega.freq+1$ }} |

| | |
|---|---|
| Step 2: | Segmentation words |
| | For each node $\alpha_i$ of FreqList |
| | { While($\alpha_i.freq > 1$) |
| | {$\beta_j.length++;$ |
| | $\beta_j.word += \alpha_i.char;$ i++;} |
| | i=j;j++;i++} |
| Step 3: | Extract the feature strings |
| | For each node $\beta_j$ of SplitWordsList |
| | { if($\beta_j \subseteq \beta_k$) |
| | {$\beta_j.freq++;$ |
| | if($\beta_j.word == \beta_k.word$)<br>{ remove $\beta_k$ from SplitWordsList }}} |
| Step 4: | Output the feature strings |
| | TList←{Order SplitWordsList by freq and length} |
| | Return TList |

## 4.2. The Topic Feature Tracking

The Internet Web topic can be consisted of one or more specific events, the common features of the specific events determine the features of the topic [13]. The discovery and tracking of the topic feature mining the updates of the commonality and individuality among them by the analysis of the features of specific events, in order to achieve the trend toward backtracking.

In the actual calculation process, e-FTM leads the TF-IDF algorithm into topic feature discovery process. The designing concept of TF-IDF algorithm is that through the method of calculation the relative term frequency, filter out some words which present in some documents and have a high frequency. These words often don't have a clear topic feature, and don't have the practical effect in target text classification.

Using the filter common features effectively, e-FTM assorts Web text feature topics and events. Different from TF-IDF application (use the target text after the participle as model input), e-FTM is based on the extraction and analysis of multi-domain text feature, and make the characteristics of the target word text string as input conditions. Thought the method of TF-IDF relative frequency, e-FTM can assort the common Characteristics of the test sample set (topic features) and personality characteristics (the event features).

$\alpha$ is set as a string of feature words in target text T, $f_\alpha$ and $f_{\alpha,i}$ represent the frequency of $\alpha$ and the frequency of the sample set in the i-th paper, N represent the total sample set for the test, n is the number of samples that include $\alpha$ , $Th$ is the preset threshold value. The specific calculating method is shown in formula -1:

$$(1)\qquad \frac{f_\alpha}{\sum_{i=1}^{N} f_{\alpha,i}} \times \log(N/n) = \begin{cases} TopicFeature & (less\ than\ Th) \\ EventFeature & (greater\ than\ Th) \end{cases}$$

With TF-IDF algorithm, e-FTM effectively classifies the common and the personality feature word of the target sample set, establishes domain topics and domain events sample set initially. In order to track the topic event and its development trend, the model divided domain event feature sample set determined event feature sample set and candidate events feature sample set in the calculating process, to store the events that are certain to be done and the events that will be done respectively. When the candidate events meet the certain frequency, it will be migrated to determined event.

 

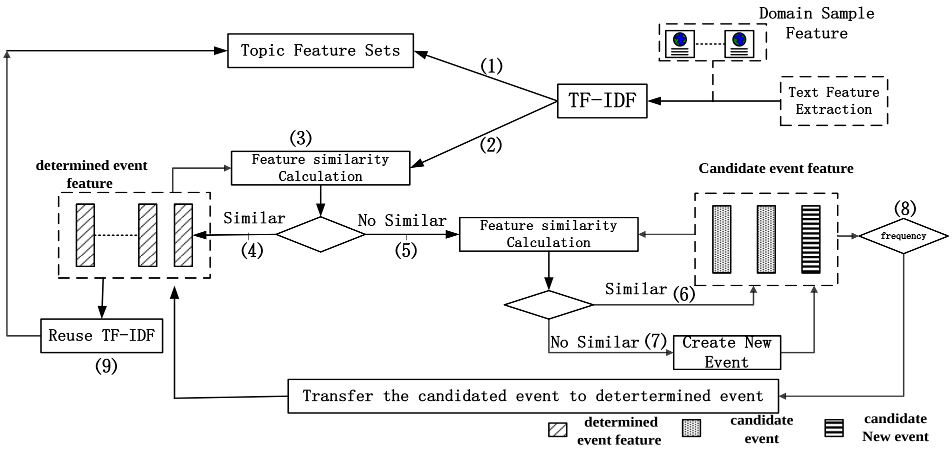


**Figure 2. The Process of Topic Feature Discovery and Tracking**

**Table 4. The Process Description of Topic Feature Discovery and Tracking**

| No. | Process Content | No. | Process Content |
|---|---|---|---|
| ① | Topic feature separation | ② | Event feature separation |
| ③ | The similarity calculating of target event feature and determined event feature | ④ | If there is a determined event feature that is similar to the target event feature, update the frequency of the determined event and its feature strings |
| ⑤ | If there is not a determined event feature that is similar to the target event feature. The model calculates if there is a candidate event that is similar to it | ⑥ | If there is a candidate event feature that is similar to the target event feature, update the frequency of the candidate event and its feature strings |
| ⑦ | If there is not any candidate event feature that is similar to the target event feature, then create the new candidate event, and record its feature strings | ⑧ | When the candidate events meet the certain frequency, it will be migrated to determined event |
| ⑨ | Multiplex TF-IDF algorithm to extract the common features of determined events, and add it to the topic feature set | | |

Combine with the description in Table 4, set $\alpha_{TList}$ as the determined event feature vector, set $\beta_{TList}$ as the target event feature vector, set $\theta_{TList}$ as the candidate event feature vector. Among them, $\omega$ , $\psi$ , $\gamma$ represent the feature string of the above vectors. $New_{TList}$ represents the new event that may be found, $\varphi$ is one feature string of $New_{TList}$, discuss the following three conditions respectively.

when $\exists\, \alpha_{TList}$ is similar to $\beta_{TList}$, *i.e.*, there is a certain amount of $\omega.word == \psi.word$ , then update all the frequency of the feature string that meet the conditions （ $\omega.freq = \omega.freq + \psi.freq$ ）, the determined event $\alpha_{TList}$ plus 1 （ $\alpha_{TList}.freq{+}{+}$ ）.

when $\forall \alpha_{TList}$ is not similar to $\beta_{TList}$, but $\exists \theta_{TList}$ is similar to $\beta_{TList}$, then update the frequency of the candidate event $\beta_{TList}$ and the feature string $\gamma$.

when $\gamma.freq = \gamma.freq + \psi.freq$ en $\forall \alpha_{TList}$ is not similar to $\beta_{TList}$, and $\exists \theta_{TList}$ is not similar to $\beta_{TList}$, then create the new candidate event vector $New_{TList}$, and initialize it as （ $\varphi.freq = \psi.freq$ , $\varphi.word = \psi.word$ , $New_{TList}.freq = 1$ ）.

In the constant update process of candidate events feature set, when there is a frequency of the candidate event greater than the threshold（ $\exists \theta_{TList}.freq \geq Th$ ）, this vector is migrated to determined event vector $\theta_{TList} \rightarrow \alpha_{TList}$. With the increase of the determined event, in order to guarantee real-time of the detection and tracking of topic feature, e-FTM use the TF-IDF algorithm to secondary mining common features of the determined event feature set, and this powerfully improved the ability to learn of the model feature topic.

## 5. Experimental Verification

### 5.1. Experiment Design

In order to verify the performance, we use web crawler technology to collect more than 150000 academic paper abstracts from Chinese full-text journals database between 05 to 08 including biological, medical, machinery, materials during the experiment, establish the Multi-domain Academic Topic Feature Testing Corpus(MATF for short). The experiment verifies practical effect of the model by the verification of multi-domain feature extraction, topic features dynamic tracking and domain knowledge collaborative scheduling. Detailed information of MATF is shown in Table 6.

**Table 6. Information of MATF Corpus**

| Domain | Number of topics | Number of events | Number of total sample |
|---|---|---|---|
| Biological | 4 | 64 | 50444 |
| Medical | 3 | 52 | 21992 |
| Machinery | 3 | 61 | 23088 |
| Material | 4 | 75 | 55304 |

### 5.2. Experiment Results

In the experiments, the model use 600 academic paper abstracts among the four different areas (biological, medical, machinery, and materials) as the Random sampling of the model input. Evaluate the effectiveness of the feature extraction model by the comparison and analysis in absolute word frequency statistics model (short for WF) and the TF-IDF model in Precision, Recall and F－measure. Of the above two models both use JE to realize Chinese text segmentation.

The experimental results: e-FTM put the technology of word segmentation with no dictionary into the discovery process of web text features, improve the ability to recognize the multi-domain feature, it has certain practical value.

 

**Table 7. The Comparison Testing of Text Feature Extraction**

| Methods | Precision | Recall | F-measure |
|---|---|---|---|
| WF | 49.3% | 37.2% | 42.40% |
| TF-IDF | 61.2% | 58.6% | 59.87% |
| e- FTM | 74.2% | 71.5% | 72.82% |

As shown in Table 7, WF and TF-IDF are two kinds of model of feature extraction based on Chinese word segmentation. The model is simple and easy to operate, but due to the limitation of word segmentation dictionary, they cannot effectively realize the word segmentation of multi-domain words, and this seriously affect the accuracy of the model (the accuracy of WF is 49.3%, the accuracy of TF-IDF is 61.2%). The introduction of the word segmentation technology without dictionary, make the e-FTM get rid of the reliance of the domain dictionary, improve the adaptability and portability of the model in multi-domain environment (the accuracy of Precision is 74.2%, the accuracy of Recall is 71.5%).

## 6. Conclusions and Future Work

Traditional domain-oriented information extraction models are always based on domain dictionary or marked thesaurus for mining text feature. Aiming at this shortage, this paper put forward a multi-domain web text feature extraction model for e-Science environment. This model put the technology of word segmentation without dictionary into the discovery process of web text features, and effectively gets rid of the dependence that traditional methods for domain dictionary, to some extent, enhance the portability and practical value of the model in multi-domain science date. By assorting the domain topics and its inherent characteristics of specific events, the model extracts the characteristics of d events time to time and excavates characteristic vocabulary in different periods, so it can achieve the track about the occurrence and development trend of domain events, gradually form a plurality of regional data centres in many domains, and provide a good data prototype to meet a variety of application needs of researchers. To improve the utilization efficiency of domain knowledge within the scope of global, the model use cases to represent high-frequency events of each data centres abstractly. With cooperative scheduling for cases of domain event in each data centre, the model can achieve collaboration and sharing of information in many domains. In the verification experiment, we collected scientific literature summary of Chinese full-text periodical database as a test corpus, and get on prototyping in three aspects, that text feature extraction in multi-domain (accuracy rate 74.2%, recall rate 71.5%,F-index 72.82%), topics feature tracking and domain knowledge cooperative scheduling, and then proved the reasonableness and practical value of this technology.

Future research work, we will continue to combine the technology of word segmentation based on the practical application of multi-domain environments demand, to design the better software architecture, and increase extraction and discovery capabilities in multi-domain text feature gradually. On the other hand, the model will be realized by the MapReduce Technology, to improve its expansion capability in large-scale massive data environment.

## References

[1] K. Khelif, R. Dieng-Kuntz and P. Barbry, “An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain”, Journal of Universal Computer Science, vol. 13, no. 12, **(2007)**, pp. 1881-1907.

[2] T. McIntosh and J. R. Curran, “Challenges for extracting biomedical knowledge from full text”, Biological, translational, and clinical language processing, **(2007)**, pp. 171-178.

[3] Z. Ozkan Gokturk, N. Kesim Cicekli and I. Cicekli, "Metadata Extraction from Text in Soccer Domain", 23rd International Symposium on Computer and Information Sciences, **(2008)**, pp. 1-6.
[4] R. D. Nielsen, W. Ward, J. H. Martin and M. Palmer, "Extracting a Representation from Text for Semantic Analysis", Proceedings of ACL-08, **(2008)**, pp. 241-244.
[5] V. Dahl and B. Gu, "Semantic Property Grammars for Knowledge Extraction from Biomedical Text", ICLP 2006, LNCS 2006, 4079, pp. 442-443.
[6] B. Martins, H. Manguinhas, J. Borbinha and W. Siabato, "A geo-temporal information extraction service for processing descriptive metadata in digital libraries", e-Perimetron, vol. 4, no. 1, **(2009)**, pp. 25-37.
[7] H. Guo, H. Zhu and Z. Guo, "Domain Adaptation with Latent Semantic Association for Named Entity Recognition", The 2009 Annual Conference of the North American Chapter of the ACL, **(2009)**, pp. 281-289.
[8] R. Verma, "Extraction and Classification of Emotions for Business Research", ICISTM 2009, **(2009)**, pp. 47-53.
[9] V. Calais Pedro, R. Stefan Niculescu and L. Vlad Lita, "Okinet: Automatic Extraction of a Medical Ontology from Wikipedia", Proceedings of WiKiAI08, **(2008)**, pp. 37-42.
[10] J. Mori, M. Ishizuka and Y. Matsuo, "Extracting keyphrases to represent relations in social networks from web", Proceedings of the Twentieth IJCAI Conference, **(2007)**, pp. 2820-2825.
[11] S. Van Landeghem, Y. Saeys and Y. Van de Peer, "Extracting Protein-Protein Interactions from Text using Rich Feature Vectors and Feature Selection", Proceedings of Third International Symposium on Semantic Mining in Biomedicine, **(2008)**, pp. 77-84.
[12] E-Science. http://en.wikipedia.org/wiki/E-Science.
[13] L. Xiao-Guang, Y. Ge, W. Da-Ling and B. Yu-Bin, "Latent Concept Extraction and Text Clustering Based on Information Theory", Journal of Software, vol. 19, no. 9, **(2008)**, pp. 2276-2284.
[14] tf–idf. http://en.wikipedia.org/wiki/Tf–idf. **(2014)** November.
[15] RDF Primer, http://www.w3.org/TR/rdf-primer, **(2014)** November.
[16] H. De-cai, Q. Hua-chun and Q. Neng, "TS-PageRank Algorithm Based on the Model of Topic Similarity", Journal of Chinese Computer Systems, vol. 28, no. 3, **(2007)**, pp. 510-514.

## Author

**Weng Yu**, is associate Professor of School of Information Engineering, Minzu University of China, Beijing.

His current research interests are distributed computing, cloud computing, and network public opinion analysis.