Sādhanā Vol. 39, Part 1, February 2014, pp. 53–62. © Indian Academy of Sciences

Mining knowledge from text repositories using information extraction: A review

SANDEEP R SIRSAT[1,*], DR VINAY CHAVAN[2]
and DR SHRINIVAS P DESHPANDE[3]

[1]Department of Computer Science, Shri Shivaji Science and Arts College,
Chikhli 443 201, India
[2]Department of Computer Science, S K Porwal College, Kamptee, Nagpur 441 002, India
[3]P G Department of Computer Science and Technology, Degree College of Physical Education, Hanuman Vyayam Prasarak Mandal, Amravati 444 605, India
e-mail: sandeep_sirsat@rediff.com; drvinaychavan@yahoo.co.in;
shrinivasdeshpande68@gmail.com

MS received 24 August 2012; revised 25 April 2013; accepted 13 August 2013

Abstract. There are two approaches to mining text form online repositories. First, when the knowledge to be discovered is expressed directly in the documents to be mined, Information Extraction (IE) alone can serve as an effective tool for such text mining. Second, when the documents contain concrete data in unstructured form rather than abstract knowledge, Information Extraction (IE) can be used to first transform the unstructured data in the document corpus into a structured database, and then use some state-of-the-art data mining algorithms/tools to identify abstract patterns in this extracted data. This paper presents the review of several methods related to these two approaches.

Keywords. Information extraction (IE); text mining; text repositories; knowledge discovery from database (KDD).

1. Introduction

In the contemporary era, most of the information is available in the form of unstructured natural language documents due to the growth of the web, digital libraries, technical documentation, etc. It is the need of the time to discover non-trivial, previously unknown, and potentially useful knowledge from such unstructured natural language documents. Hence discovering useful knowledge from unstructured text, i.e., text mining, is becoming an increasingly important aspect of Knowledge Discovery (Mahgoub *et al* 2007; Mooney and Nahm 2003; Clifton *et al* 2004; Ding *et al* 2011; Fatudimu *et al* 2008). Thus text mining is an increasingly important research

[*]For correspondence

field, and is similar in some sense with data mining, as both are included in the field of information mining. However, data mining deals with structured database. Thus it is important to transform the unstructured text into structure database, so that the data mining techniques may be adapted in a straightforward way to mine text.

Information extraction (IE) methods, with reasonable accuracy, are able to transform the unstructured text into structured database, called intermediate forms. The most usual intermediate forms are: bag-of-words, N-grams, keyphrases, multi-term phrases, concept word, concept hierarchy, conceptual graph, etc.

Most of the text mining techniques treat documents as an unordered bag-of-words. Then it specifies a weighted frequency for each of these terms in the documents as a parse vector using standard vector space model. Such a simplified representation of text has been shown to be quite effective for the number of standard tasks, such as, information retrieval (IR), classification, and clustering (Zakzouk and Mathkour 2011; Mayor and Pant 2012).

However, mining knowledge from text cannot be discovered using simple bag-of-words representation. It is not useful to assert the properties and relationship among the important entities using standard vector-space model. Thus existing methods in IE, with reasonable accuracy, are able to identify several types of entities in text documents and establish some relationships that are asserted between them (Mooney and Nahm 2003; Califf and Mooney 1997; Clifton *et al* 2004; Carlson *et al* 2010; Patwardhan and Riloff 2006). Recently developed text mining techniques describe the integration of IR methods with data mining techniques for association rule discovery (Mahgoub *et al* 2007; Fatudimu *et al* 2008).

Therefore, IE can serve as an important technology for text mining. If the knowledge to be discovered is expressed directly in the documents to be mined (Ding *et al* 2011; Carlson *et al* 2010), IE alone can serve as an effective approach to text mining. However, if the documents contain concrete data in unstructured form rather than abstract knowledge, IE can be used to first transform the unstructured data in the document corpus into a structured database, and then use some state-of-the-art data mining algorithms/tools to identify abstract patterns in this extracted data (Mahgoub *et al* 2007; Mooney and Nahm 2003; Clifton *et al* 2004; Fatudimu *et al* 2008).

2. Information extraction: Problems and methods

Information Extraction (IE) is concerned with locating specific set of relevant items from natural language documents. Thus, IE systems can extract structured information from unstructured text. One type of IE is *named entity extraction* and then creation of filled templates (Konchady 2009). The named entity extractor identifies references to particular kinds of objects such as names of people, companies, and locations. Duan & Zheng (2011) studied the features of the Chinese Named Entity Recognition (CNER) based on conditional Random Fields (CRFs). These features include common attributes, feature templates varying in windows size (3, 5, 7) and sequence labels sets. In addition to recognize entities, it is useful to specify specific types of relations between entities. The KnowITALL system is able to extract instances of relations, such as capitalOF (City, Country), or starsIN (Actor, Film) (Etzioni *et al* 2005). Doug Downey *et al* (2002) applied a simple pattern learning algorithm, to the task of IE that can be used as both extractors (to extract the instances of relations) and discriminators (to access the truth of extracted information). Callan and Mitamura (2002) presented a new approach to named-entity detection, known as KENE, which uses knowledge-based approach for learning extraction rules. It uses generate-and-test approach to named-entity extraction from structured documents. Carlson *et al* (2010) consider the problem of semi-supervised learning to extract categories (e.g., academic fields, athletes) and relations e.g., PlaySport (athlete, sport) from web pages, starting with a handful labelled seed example for each, and a set of constraints that couple the various categories and relations. This approach shows that, training both contextual pattern extractor that extract information from freeform text (e.g., the pattern 'Mayor of arg1' as an extractor for the category <city>) and wrapper which extracts information from semi-structured documents (e.g., the wrapper "<td Class = "City">arg1</td>" from some specific URL).

It can also be used to extract fillers for a predefined set of slots (rules) in a particular template (frame) relevant to the domain. RAPIER developed by Califf and Mooney (1997) consider the task of extracting database from posting to the USENET newsgroup, *austin.Job*. This system uses an automatic pattern based learning approach that extracts rules for identifying each type of entity or relation. The learned extraction rules consist of three parts: (i) A pre-filler pattern that matches the text immediately preceding the phrase to be extracted. (ii) A filler pattern that matches the phrase to be extracted, and (iii) a post filler pattern that must match the text immediately following the filler. It uses regular expression (regexs) languages similar to that used in Perl (Billisoly 2008), to express the pattern that utilizes limited syntactic information, produced by POS tagger (Konchady 2009). It consists of a specific to general (bottom-up) search for pattern that characterizes slot fillers and their surrounding context.

Another approach of IE is to extract structured data (pattern) from unstructured or semi-structured web pages. Pattern induction to generate extraction patterns from a number of training instances is one of the most widely applied approaches to IE. Kim *et al* (2009) proposed a local tree alignment based soft pattern matching method for IE. This method considers the node labels, as well as link labels to the head node, because the class of link to the head node is important as the node label itself for dependency trees. Moreover, the method also considers the alignment of slot value nodes in the tree patterns for adapting information extraction task. If the pattern node v is a kind of slot value nodes, the similarity score between v and w is inherited from parents of both nodes. It then constructs the pattern candidate sets for four types of pattern representation models, based on the dependency trees and scenario templates of the training data. For each pattern candidate, corresponding confidence score and optional threshold value were computed and arranged in descending order of confidence score.

Another general approach to IE is to treat it as a sequence labelling task in which each word (token) in the document is assigned a label (tag) from a fixed set of alternatives. One approach to the resulting sequence labelling problem is to use a statistical sequence model, such as hidden Markova model (HMM) (Konchady 2009), or a conditional Random field (CFR) (Duan and Zheng 2011).

3. Extracting knowledge from text

If the information extracted from a corpus of documents represents abstract knowledge rather than concrete data, IE itself can serve as 'discovering' knowledge from text. Discovery of knowledge by extracting information, such as, keyphrases/keywords extraction from text found useful for many other text mining tasks, such as, classification, clustering, summarization, topic detection, etc.

One of the approaches to extract relevant information from the related topic is the selection of one or more phrases that best represent the content of the documents. Unlike the IE task, it does not consider any known fields (slots) for a template. Instead, text segments that are unique and most representative of documents are extracted. Most systems use TF-IDF scores to sort the phrases in multiple text documents and select the top-k as keyphrases. Many existing methods convert the keyphrase extraction as classification problem using S V M (Zakzouk and Mathkour 2011; Mayor and Pant 2012). Ding *et al* (2011) proposed a novel formulation, which present several criteria of high quality news keyphrase and integrate those criteria into the keyphrase extraction task by converting the task to the binary integer programming (BIP) problem. In this approach the proposed BIP based method can combine the unsupervised methods, such as, TF-IDF and locality information, as assignment value in the object function. This method considers several constraints converted from the coverage and coherence criteria, and the number of extracted phrases. It assumes that high quality keyphrases should cover the whole document or group of documents in a right order. First, to satisfy coverage criteria, the Latent Dirichlet Allocation (LDA) model is used to estimate words distribution over topic. Second, it uses mutual information (MI) to measure the word coherence, which should satisfy other criteria that the keyphrases should be semantically related and coherent. In this case, the keyphrases pair with high occurrence frequency are selected together. An experimental result proved that TF-IDF is the most important feature and locality feature can further improve the performance.

Another approach presented by Bhattacharya *et al* (2010), focuses on dictionary-based text mining and its role in enabling practitioners in understanding and analysing large text datasets. To build a concept dictionaries for annotating a collection of documents from a particular domain, they define dictionary $\mathrm{D} = \mathit{Dict}(\mathrm{C}, \mathrm{X})$ as a set of words that refers to or describes a semantic concept C in a document collection X. In this method, an online interactive framework is applied, where the user starts off with a small set of words, inspects the results, selects and rejects words from the returned ranking, and iterates until get satisfied. With interactive supervision, the user provides positive and negative seed words at each stage of iteration to the algorithm. This process gradually refines the seed sets and the ranking comes closer to the user's preference as the iteration continues.

This framework of constructing dictionary needs to provide a set of seed words for specifying a concept C and refers the WordNet to define the semantics of seed set unambiguously, for general purpose English words. However, ambiguities may arise in selecting the words in seed set, or some subset of them, when all words are not identical to these concepts or the conceptual structure is absent like WordNet. Further, the dictionaries need to be constructed for every new dataset and the existing concept nodes can be used for seeding. Then the ranking returned by the system is inspected to create the adapted dictionary for the new document collection. Thus re-using dictionaries can significantly make easier the task of specifying the semantic concept in the absence of semantic structure for dictionary construction for a concept.

The observation made from the experimental results suggest that interactively building dictionaries from scratch leads to good dictionaries, but adapting earlier dictionaries also leads to dictionaries of similar quality and adapting dictionaries consistently results in 50–60% time savings. One shortcomings of this supervise model is its need to provide the good quality positive and negative set of seed words. Again, the system is unable to measure benefits in terms of precision and recall, as extensive experimentation is required due to lack of public tagged corpora.

4. Mining knowledge from text

In many cases, the information extracted from unstructured text represents specific data rather than abstract knowledge. In such situation, the text mining task requires to perform some

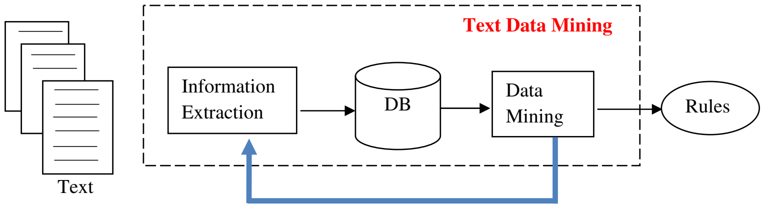

Figure 1. Overview of IE-based text mining framework.

additional process to mine knowledge from this specific data. Mooney and Nahm (2003), Clifton *et al* (2004). One approach to text mining is to first use IE to obtain structured data from unstructured text and then use traditional Knowledge Discovery from Database (KDD) tools to discover knowledge from this data. Mooney and Nahm (2003) present a framework for text mining, called DiscoTEX (Discovery from Text EXtraction) as shown in figure 1. It uses learned IE system to transform unstructured text into more structured data and then mine this data to form interesting relationship.

The IE learning system of DISCOTEX integrates IE module acquired by an IE learning system, and a standard rule induction module. The IE module used two state-of-the-art system for learning information extractors, RAPIER (Robust Automated Production of Information Extraction Rules) (Califf and Mooney 1997) and BWI (Boosted Wrapper Induction). The RAPIER system performed well on realistic applications such as USENET job posting. After constructing an IE system that extracts the desired set of slots for a given application, IE extraction patterns is applied to each document of a text corpus to create a collection of structured records (database). Standard KDD techniques can then be applied to the resulting database to discover interesting relationship. Sample rules mined from a database of 600 resumes extracted from the USENET newsgroup *misc.jobs.resumes* by BWI are shown in table 1. Specially, DISCOTEX induce rules for predicting each piece of information in each database field given all other information in a record. To discover prediction rule, each slot value pair in the extracted database is treated as a distinct binary feature.

It then applies C4.5 RULES to discover interesting rules from the resulting binary data. Sample rules that C4.5rules mined from a database of 600 jobs that RAPIER extracted from the USENET newsgroup *austin.jobs* are shown in table 2. Discovered knowledge describing the relationship between slot values is written in the form of production rules. For example, if there is a tendency for “Web” to appear in the ‘area’ slot, when “Director” appears in the ‘application’ slot, this is represented by the production rule.

$$\text{‘Director} \in \text{application} \rightarrow \text{Web} \in \text{area’}$$

The major drawback of DISCOTEX is that, it focuses on rules predicting the presence of fillers rather than predicting the absence of filler in a slot.

Table 1. Sample rules mined from CS resumes.

- HTML ∈ language and DHTML ∈ language → XML ∈ language
- Dreamweaver 4 ∈ application and Web Design ∈ area → Photoshop 6 ∈ application
- ODBC ∈ application → JSP ∈ language
- Perl ∈ language and HTML ∈ language → Linux ∈ platform

Table 2. Sample rules mined from CS job postings.

- Oracle $\in$ application and QA Partner $\in$ application $\rightarrow$ SQL $\in$ language
- Java $\in$ language and ActiveX $\in$ area and Graphics $\in$ area $\rightarrow$ Web $\in$ area
- $\neg$(UNIX $\in$ platform) and $\neg$(Windows $\in$ platform) and Games $\in$ area $\rightarrow$ 3D $\in$ area
- AIX $\in$ platform and $\neg$(Sybase $\in$ application) and DB2 $\in$ application $\rightarrow$ Lotus Notes $\in$ application

Clifton *et al* (2004) developed another technique TopCat (Topic Categories) for identifying topics that recur in articles of text corpus. This method used IE to identify named entities in individual articles and represent them as a set of items of an article. Thus they view the problem in data mining/database context, by identifying frequent itemsets that is group of named entities that commonly occurred together. TopCat use association rule data mining technique for identifying these frequent itemset. It further clusters the named entities, using a hypergraph splitting technique, which finds, group of frequent itemsets with considerable overlap. Then it applied IR technique to find document related to the topic. This approach uses disparate technologies, such as, IE for named entity extraction, association rule data mining, clustering of association rules, IR techniques, and few specific development that have wider applications. TopCat identifies topics with reasonable accuracy with understandable identifiers for the topic. For example, the most important three topics identified based on the support of the frequent itemset used in generating the topics, are shown in table 3.

The observation from experimental result leads to the conclusion that:

- Evaluating TopCat is difficult.
- TopCat is relatively insensitive to errors in named entity tagging.
- The segmentation of stories is more critical — since the result produced by TopCat is unreliable, if many documents contain multiple unrelated stories.

Table 3. Top three topics for January through June 1998.

Topic 1		Topic 2		Topic 3	
Location	Baghdad	Location	Alaska	Location	Albania
Location	Britain	Location	Anchorage	Location	Macedonia
Location	China	Location	Caribbean	Location	Belgrade
Location	Iraq	Location	Great Lakes	Location	Bosnia
Org.	Security Council	Location	Gulf Coast	Location	Pristina
Org.	United Nations	Location	Hawaii	Location	Yugoslavia
Person	Kofi Annan	Location	New England	Location	Serbia
Person	Saddam Hussein	Location	Northeast	Person	Slobodan Milosevic
Person	Richard Butler	Location	Northwest	Person	Ibrahim Rugova
Person	Bill Richardson	Location	Ohio Valley	Org.	Nato
Location	Russia	Location	Pacific Northwest	Org.	Kosovo Liberation Army
Location	Kuwait	Location	Plains		
Location	France	Location	Southeast		
Org.	U N	Location	West		
		Person	Byron Miranda		
		Person	Karen Mcginnis		
		Person	Meteorologist Dave Hennen		
		Person	Valerie Voss		

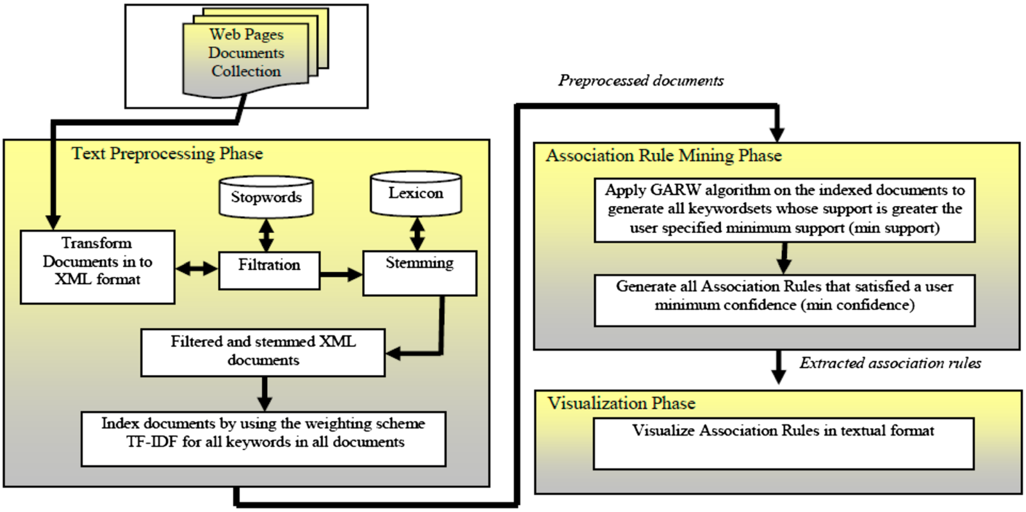
Web Pages
Documents
Collection
Preprocessed documents
Text Preprocessing Phase
Stopwords
Lexicon
Transform
Documents in to
XML format
Filtration
Stemming
Filtered and stemmed XML
documents
Index documents by using the weighting scheme
TF-IDF for all keywords in all documents
Association Rule Mining Phase
Apply GARW algorithm on the indexed documents to
generate all keywordsets whose support is greater the
user specified minimum support (min support)
Generate all Association Rules that satisfied a user
minimum confidence (min confidence)
Extracted association rules
Visualization Phase
Visualize Association Rules in textual format

Figure 2. Text mining system architecture.

It raises two issues: (i) Is it possible to incrementally update new topics without looking at the old data? (ii) How to make alert the user when new knowledge is being added to the topic?

An alternative approach to discover knowledge from text is to combine the information retrieval (IR) scheme (TF-IDF) for keyword/feature selection with association rule data mining technique for discovering knowledge (rules). Mahgoub *et al* (2007) presented new text mining technique called, Extracting Association rules from text (EART), for automatically extracting association rules from collection of text documents, the architecture of EART system is shown in figure 2.

The EART system integrates the XML technology with IR scheme TF-IDF for selecting the most discriminative keywords/features. The frequency based weighting scheme TF-IDF (Mahgoub *et al* 2007; Clifton *et al* 2004; Fatudimu *et al* 2008; Konchady 2009) is used to index the documents by assigning higher weights to distinguished term in a document. The system then sort out the keyword based on their weight score and selects only the top N frequent keywords up to M% of the number of running words. It then applies the own designed algorithm for

Table 4. Example of Medline abstract (abbreviated).

Text: # 24
Title: Investigation of avian influenza outbreak in humans
Author: Kash JC, Goodman AG, Korth MJ, Katze MG
Abstract: Recently, there is much concern over the highly pathogenic avian influenza H5N1 viruses into the human population. Influenza virus has evolved complex translational control strategies that utilize cap-dependent translation initiation mechanisms and involve the recruitment of both viral and host-cell proteins to preferentially synthesize viral proteins and prevent activation of antiviral responses.
Infected birds have been the primary source of influenza infection in humans in Asia.
But the transmission from poultry to humans is very limited at present, and requires a direct exposure to live birds, whereas there was no significant risk related to eating well-cooked poultry meat. . .

Table 5. A sample of the generated association rules.

Highly, pathogenic, avian influenza –> H5N1	80%
Outbreak, H5N1, poultry –> Asia	72%
Viruses, H5N1, isolate –> Vietnam	69%
Avian influenza, virus, poultry –> human	70%
Outbreak, avian influenza, Thailand –> Vietnam	85%
Epidemic, H5N1, avian influenza –> Asia	83%
Infect, humans, Asia –> influenza	83%
Pandemic, virus –> human, transmission	60%
Virus, isolate –> influenza, H5N1	50%
Spread –> highly, pathogenic, avian influenza	67%

Generating Association Rules based on Weighting scheme (GARW) for discovering most useful association rules.

The excerpt of document consisting medline abstract is as shown in table 4, and the samples of generated association rules is shown in table 5.

The rule, such as,

"highly, pathogenic, avian influenza –> *H5N1* 80%"

tells that in 80% of texts, where the three words (highly, pathogenic, avian influenza) occurred within 3 consequent words, the word H5N1 co-occurs within 4 words. Similarly, the rule

"*outbreak, H5N1, poultry –> Asia* 72%"

tells that in 72% of the texts, the three words (*outbreak, H5N1, poultry*) occurred within 3 consequent words, the word *Asia* co-occurs within 4 words.

Fatudimu *et al* (2008) developed a system similar to EART that is applied to per-election information collected from the website of the Nigerian Guardian newspaper. The extracted association rules contain important features and describe the informative news included in the documents collection related to the concluded 2007 presidential election. The useful information presented by system could help to sanitize policy as well as to protect the nascent democracy.

5. Conclusion

In this paper, we have reviewed several methods related to two approaches for mining text from text repositories using IE. In the first one, IE is used to extract abstract knowledge from text rather than concrete data. Regarding the first approach we discussed two methods in brief. One is related to keyphrase extraction, whereas the other focuses on building concept dictionaries for annotating a collection of documents from a particular domain.

In the second approach, IE is used to obtain structured data from unstructured text and then KDD is applied to discover knowledge from this structured data. We mainly discussed two systems, viz., DISCOTEX and TOPCAT. Both use IE to view the problem in data mining/database context and then association rule mining technique is applied to discover the useful knowledge (rules). However, we also discussed other systems, EART and other similar one, that combines IR scheme TF-IDF (for keyword/feature extraction) and association rule mining techniques for discovering knowledge. The key feature of EART system is that, the extracted association rules get the relations among the existing keywords in text documents collection ignoring the order in which these keywords occur. It has been proved by the experimental observation that these techniques help to improve the performance of the system by reducing execution time and extracts more interesting and useful rules than the rules generated by other systems. The experimental result of paper Ding *et al* (2011) also proved that the contribution of TF-IDF component in objective function enhance the result by improving precision from 58.86% to 71.45%, recall from 60.82% to 73.96% and F-measure from 59.82% to 72.68%.

In conclusion, authors would like to state that Most IE systems are developed using supervised approach. They either use human annotated corpora or integrative framework to train the system. However, constructing sufficient corpus for training accurate IE system is tedious and time consuming task. Further, interactive framework based system requires to provide good quality set of seed words to the system. One approach is to use automatic learning methods to decrease the amount of training data that uses vast repositories of annotated text. Patwardhan & Riloff (2006), presented an approach to exploit an existing set of IE patterns that were learned from small set of annotated training data to automatically identify new, domain-specific text from the web. These web pages are then used for additional IE training, yielding a new set of domain specific IE patterns. However, more research is needed to explore methods for reducing the demand for supervised training data in IE.

Another approach to reduce the demand for supervised corpus-based training system is to develop unsupervised learning methods for building IE system. However, some work has been done in this area Clifton *et al* (2004), Etzioni *et al* (2005). This is another promising area for future research.

One of the most promising and emerging techniques in discovering knowledge from text is to combine the IR schemes (such as TF-IDF) and data mining techniques (such as association rule mining) to discover knowledge (rules). We discussed some approaches Mahgoub *et al* (2007), Clifton *et al* (2004), Fatudimu *et al* (2008) in brief in Section 4. This approach can be further extended to use the concept feature to represent the text and to extract more usefulness that is more meaningful to represent the contents of the text document.

Another alternative approach to text data mining (TDM) is text knowledge mining (TKM) (Sánchez *et al* 2008; Davi de Castro *et al* 2004). One of the approaches presented by Davi de Castro *et al* (2004) is based on the concept of tree-edit distance that allows not only the extraction of relevant text passages from the page of a given website, but also the identification of pages of interest and the extraction of a relevant text passages discovering the non-useful material.

References

Bhattacharya I, Godbole S and Gupta A 2010 Building re-usable dictionary repositories for real-world text mining - *CIKM'10*, October 26–30, 2010, Toronto, Ontario, Canada

Billisoly R 2008 Practical text mining with Perl, John Willey & Sons, Inc., Hoboken, New Jersey

Califf M E and Mooney R J 1997 Relational learning of pattern match rules for information extraction. In: T M Ellison (ed.) CoNLL97: Computational Natural Language Learning, ACL, pp 9–15

Callan J and Mitamura T 2002 Knowledge based extraction of named entities - *CIKM'02*, pp 532–537, November 4–9, McLean, Virginia, USA, ACM New York, NY, USA

Carlson A, Betteridge J and Wang R C 2010 Coupled semi-supervised learning for information extraction-WSDM'10, February 4–6, New York City, New York, USA

Clifton C, Cooley R and Rennie J 2004 TopCat: Data mining for topic identification in a text corpus. *IEEE Trans. Knowl. Data Eng.* 16(8): 949–964

Davi de Castro R, Golgher P B, da Silva A S and Laender A H F 2004 Automatic web news extraction using tree edit distance-WWW2004, pp 502–522, May 17–22. ACM, New York, USA

Ding Z, Zhang Q and Huang X 2011 Keyphrase extraction from online news using binary integer programming, *Proc. 5th Internat. Joint Conf. on Natural Language Processing*, Chiang Mai, Thailand, pp 165–173, November 8–13

Downey D, Etzioni O, Soderland S and Weld D S 2002 Learning text patterns for web information extraction and assessment (www.aaai.org)

Duan H and Zheng Y 2011 A study on features of the CRFs-based Chinese named entity recognition. *Int. J. Adv. Intell.* 3(2): 287–294

Etzioni O, Cafarella M, Downey D, Popescu A-M, Shaked T, Soderland S, Weld D S and Yates A 2005 Unsupervised named-entity extraction from the web: An experimental study. *Artif. Intell.* 165(1): 1–42

Fatudimu I T, Musa A G, Ayo C K and Sofoluwe A B 2008 Knowledge discovery in online repositories: a text mining approach. *Eur. J. Sci. Res.* 22(2): 241–250

Gupta V and Lehal G S 2009 A survey of text mining techniques and applications. *J. Emerg. Technol. Web Intell.* 1(1): 60–76

Kim S, Jeong M and Lee G G 2009 A local tree alignment-based soft pattern matching approach for information extraction: *Proc. of NAACL HLT 2009: Short Papers*, Boulder, Colorado, pp 169–172, June, Association for Computational Linguistic

Konchady M 2009 Text mining application programming- Cengage Learning India Private Ltd.

Mahgoub H, Rosner D, Ismail N and Torkey F 2007 A text mining technique using association rules extraction. *Int. J. Comput. Intell.* 4(1): 21–28

Mayor S and Pant B 2012 Document classification using support vector machine. *Int. J. Eng. Sci. Technol.* (IJEST) 4(4): 1741–1745

Mooney R J and Nahm U Y 2003 Text mining with information extraction- multilingualism and electronic language management, *Proc. 4th Internat. MIDP Colloquium*, September 2003, Bloemfontein, South Africa, W Daelemans, T du Plessis, C Snyman and L Teck (eds) Van Schaik Pub., South Africa, pp 141–160

Patwardhan S and Riloff E 2006 Learning domain-specific information extraction patterns from the Web-IEBeyondDoc '06, *Proc. Workshop Information Extraction Beyond The Document*, *Association for Computational Linguistics*, Stroudsburg, PA, USA, pp 66–73

Rose S, Engel D, Cramer N and Cowley W 2010 Automatic keyword extraction from individual document, Text mining: Application and theory, M W Berry and J Kogan (eds) John Willey & Sons Ltd 2010, pp 3–20

Sánchez D, Martín-Bautista M J and Blanco I 2008 Text knowledge mining: an alternative to text data mining. *IEEE Int. Conf. Data Mining Workshops*, pp 664–672. doi:10.1109/ICDMW.2008.57

Shehata S, Karray F and Kamel M S 2010 An efficient concept-based mining model for enhancing text clustering. *IEEE Trans. Knowl. Data Eng.* 22(10): 1360–1371

Speretta M and Gauch S 2008 Using text mining to enrich the vocabulary of domain ontologies - 2008 *IEEE/ WIC/ ACM Internat. Conf. Web Intelligence and Intelligent Agent Technology*, vol. 1, pp 549–552, *IEEE Computer Society*, Washington, DC, USA

Zakzouk T and Mathkour H 2011 Text classifiers for cricket sports news - 2011. *Internat. Conf. Telecommun. Tech. Appli.*, *Proc. CSIT*, vol. 5, IACSIT Press, Singapore