Indian Journal of Science and Technology, *Vol 8(35), DOI: 10.17485/ijst/2015/v8i35/55134, December 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Survey of Sentiment Analysis for Journal Citation

G. Parthasarathy[1]* and D. C. Tomar[2]

[1]Department of Computer Science and Engineering, Sathyabama University, Chennai – 600119,Tamil Nadu, India; amburgps@gmail.com

[2]Department of Information Technology, Jerusalem College of Engineering, Chennai - 600100, Tamil Nadu, India; dctomar@gmail.com

Abstract

Sentiment analysis approach belongs to the family of machine learning, where the objective is to discover useful patterns stored in a database. Due to wide availability of data, there is an upcoming need for turning such an overwhelming amount of data into useful knowledge. In this paper we recommend different techniques available for high accuracy extraction of citations for academic papers and improve the performance in citation extraction by integration of two techniques. Therefore our aim is to automate the task to determine whether a context is positive or negative. The main goal of sentiment analysis lies in finding the polarity of citation expressed in different research article. In this paper we address the techniques, approaches and methods of the research which are supportive and marked as the essential field of sentiment analysis of citations in research article. This literature survey is done to study the sentiment analysis problem in journal citation to identify different trends and recommend the upcoming research directions.

Keywords: Citation Analysis, Information Retrieval, Sentiment Analysis

1. Introduction

Opinion mining or Sentiment mining involves development of a method to discover user's suggestions made in blog posts, comments, reviews or tweets, about the product, policy or a topic. Its scope is to find the attitude of a user regarding some topic. The Web is a voluminous warehouse of structured and unstructured data. The analysis of this data to dig out underlying user's opinion and sentiment is a demanding process. An important characteristic of scientific researcher literature review is the use of citations. Researchers use citations to recognize and refer to other research papers which are interrelated in some way to their individual work. Analysis of the contribution of documents implied by citations has long been a matter of curiosity for researchers. The complete biographic information relating to a cited task is the final stage of research paper in terms of citation and reference. A citation is a revelation of a task in the body of the content and contains sufficient information to distinctively find the task in the set of references[1].

There are different types of citation style in research papers:

- Textual – Syntactic citations

  Example 'Pang et al. (2002) proposed a seminal approach in sentiment classification.'
- Numbered citations

Example 'Research in the area has shown that there are several other words that invert the polarity of an opinion expressed x, where x is a number'.

A bibliographic reference (or citation) is a document having, in its bibliography, a reference to other research papers. This reference can be shown as a connection between papers, at least those in the author's mind. So the citation is a semantic feature of a document. A citation index is based on the bibliographic references having in a document, linking the document to the cited works. The citation index allows navigation backward in time (the cited documents) and forward in time (the citing documents).Thus it is a powerful tool useful for research paper retrieval[2].

*Author for correspondence

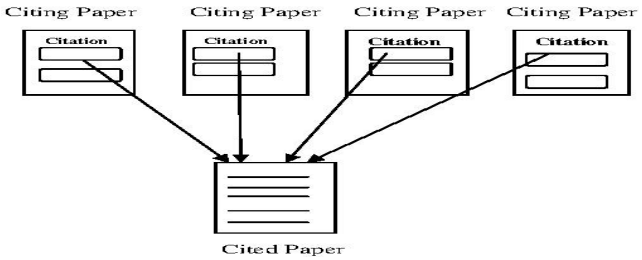

Figure 1. Research article citation relationships between other research papers.

In a research article, there are statements where the author explains the scope of a cited paper and compares the cited paper and citing paper as given above in Figure 1. The citation data are those collected from a set of research papers in that subject domain. Further, they are properly classified and ranked; they can act as a kind of a review article. In this paper, we give an overview of several methods for extract citation data from citing papers and classify the citation polarity automatically using citation links and citation style. The process of citation classification can be illustrated as follows: Given Article A and the collection of all citations S_1 to S_n citing Article A, classify every citation S_i to the suitable class C_i from the collection of classes C_1 to C_m known as a classification scheme. Every class C_i from the classification scheme specifies the purpose of the citation.

The organization of this paper is structured as follows. In Section 2, we provide journal citation sentiment analysis related work and review some of the techniques of this domain. In Section 3, describes in detail about a general view of sentiment analysis and defines the framework for sentiment mining system. In Section 4, describes citation extraction techniques and in Section 5, the feature extraction techniques are explained. The sentiment classification is explained under the Section 6, under which the polarities of citation are discussed. We analyzed the sentiment of citation and discuss the state of the art in Section 7. Finally, we conclude and describe the Future requirement in Section 8.

2. Related Literature

The essential function of the research paper from its cited frequency is to find research trends and developing areas. Due to the huge set of the online papers, some researches in citation indexing area are focused on building a universal citation database CiteSeer locates papers on the web using search engines, heuristics and web crawling. Other means of locating papers are including indexing existing archives, agreements with publishers and user submission. It also updates the citation library in regular. After the paper is located and downloaded, it is parsed to extract the citations and the context in which the citations are made. Then the citation is indexed and stored in a database. Given a research paper of request, it can also get the required papers using different measures of similarity based on term occurrence or citation information[2]. The number of research papers published in the web continues to increase and improved technology has allowed several older papers to be rapidly digitized and they are also made available in huge numbers along with that of the newly published papers.

A typically featured researcher should be able to sift through a large quantity of papers/articles manually, relying searches based on keywords or paper citations to guide them. The search results of researchers with similar interests can help to direction of this into a more effective search, but the process of sharing the search results is often too large and time consuming process to be feasible. A recommender system can help by recommending papers automatically based on the preferences of the other researchers with similar interests[3]. Researchers in currently handle these tasks by following a list of reference in previously identified papers and literature surveys, searching literature collections (using facilities such as Google Scholar, etc) and returning the existing recommendations from researchers who are already working on the topic.

Citation classification is a developing area of research that classifies citations based on the function behind the citation. To achieve citation classification, a collection of classes is required, into which citations will be classified. Besides the classes, a classifier is necessary to achieve the classification. Several classifiers have been used to achieve citation classification. It is not feasible to establish their accuracy because they use dissimilar classes. Classes described in the literature have been crafted manually by hand with different levels of granularity. Given the lack of a standard collection of classes, our aim is to close this gap by setting up a new collection of classes that we expect will be adopted as a standard by researchers in this field. The main dissimilarity between our classes and earlier introduced ones is that we do not create the classes manually; rather we use a large dataset of citations to lead us to the classes[4].

The problem of sentiment classification is the majority of binary classification problem, differentiating between *positive* and *negative* patterns. However, additional classes can also be brought in, in order to make the analysis extra robust and boost the quality of outputs. For instance, several of the tasks contain the *neutral* or *irrelevant* sentiment categories, which mean that there is no sentiment. By doing this, we can ignore the subjectivity identification process mentioned above and have the classifier differentiate between opinionative and non-opinionative phrases. There is proof for this method has a positive result on the precision of the final outputs[5,6] introduces it a syntax tree pruning and tree kernel-based approach for sentiment classification. His contribution is that utilizes the convolution kernel of SVM to obtain structured information and also applied some pruning strategies to the syntax tree for reducing the useless information.

Sentiment classification attempts to allocate the review documents either positive or negative polarity. It fails to search what the reviewer or opinion container likes or dislikes. A positive opinion on a document does not mean that the opinion holder has positive opinions on all aspects of the document. Similarly a negative opinion on a product does not necessarily mean that he/she dislikes everything about the product. In an analysis document the opinion container writes both positive and negative essence of the object, although in general opinion may be positive or negative[7].

The development of citation analysis specified in variety of citation analysis techniques is introduced using citation link and these techniques are new methods for ranking the research article[8]. Novel decision tree framework for prune the data are introduced for classification, with high accuracy and beat the limitations9. Automate the subject classification based on KSCD introduced in[13].

3. Sentiment Analysis: A General View

Sentiment analysis was initially formulated as the NLP task of retrieval of sentiments expressed in texts. A simple keyword finding will not be appropriate for mining all kinds of opinions. Hence the need for use of sophisticated opinion extraction methods. Sentiment analysis is a natural language processing technique, helps to find and dig out subjective information in source materials. Sentiment analysis aims to establish the approach of the writer with respect to the entire contextual polarity of a document. The approach may be his or her decision or valuation, sentimental state or the intended emotional communication. A fundamental process in sentiment analysis is classification of opinion polarity of a specified context at the document; whether the given opinion in an article, a context or an entity feature is positive, negative or neutral. Beyond polarity sentiment classification, the emotional states such as "irritated", "gloomy" and "joyful" are also recognized[10].

Figure 2 shows the architecture of the framework which extracts the citation from the citing articles. It consists of the following modules: Citation extraction, preprocessing the extracted citations, feature extraction, apply machine learning technique for sentiment classification and rank the result.

3.1 Framework for Sentiment Mining System

The Figure 2 illustrates the major steps for achievement of sentiment mining. The researcher wants to know the quality of research article by other citations about a particular paper. Citation data are collected from other research paper sources in such a way that only the citation is related to the paper. The citation extracted document is then preprocessed. Preprocessing, in this context, is the removal of the fact based sentences, thus selecting only the opinionated sentences. However refinement is done by eliminating the negations and by sensing the word disambiguation. Then, the task of extracting related features is finished. The extracted related features donate to a document vector upon which different sentiment or polarity classification can be applied in order to classify the polarity (positive and negative) using the obtained document vector and finally the opinion is ranked based on the citation of the researcher.

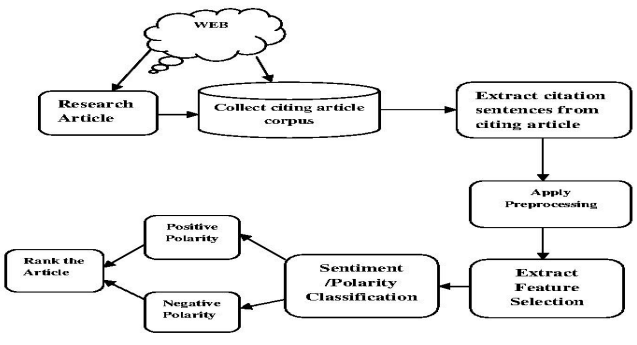

Figure 2. Architecture of sentiment mining for journal citation.

4. Citation Extraction Methods

A citation state that a well defined description of work done in the other research article, till now different techniques are developed to extract the citation. Table 1 shows the list of different citation techniques have been discussed in academic literature.

4.1 Citation Graph

The purpose of the research paper library as a graph is to dig out the citations from reference collections of papers. Then the citation graph can be built from citation links. Every vertex v in citation graph G will represent a citation document. A direct edge E_{ij} (V_i, V_j) in G will represent a link reference of V_i by V_j. The set of all the vertices in graph G is represented as V (G). The method is to form a citation graph from the extracted references. Every vertex in the graph shows a citation paper and every directed edge is a citation occurrence. Clearly the degree of the connectivity for each vertex is different. As a consequence, the highly connected sub-graphs will emerge. The in-degree of a vertex can define the importance of the citation in a research area, the out-link set of the vertex can indicate related topics that might or might not be semantically similar to the current citation. The key procedure of the approach is graph partitioning. The sub-graphs after partitioning will represent the sub-topics in the collection. The connectivity measurement can reveal the key citation, hot topics and related topics for a certain sub-graph. If partitioning procedure goes on to the finer level, the topics can be divided into the more detailed level. In this way, the subject tree of the digital library can be formed automatically and dynamically, which can solve the problem of the existing static subject tree in most of the digital libraries. So this approach can help better location of the user-desired information in many different ways.

4.2 Citation Context Free Grammar (CFG)

CFG is used to extract text from the PDF sources. The citation extraction algorithm process at the sentence level helps to isolate the tag citations. It starts with the inspection that textual citations are secured around years. To find each sentence for an author year token; if it finds such a token, our process is then to extract whether it creates part of a citation and if it does, to obtain the author names that accompany it. An example grammar for a citation is given below[3]:

<CI> :: = < AL >< W >* < YL >
<W>:: = non_author words (not accepted list)
<AL> :: = {< A_name >< A_seperator >*}
<A_seperator>:: = ,|;|and|&
<YL> :: = [(]{ year ><y_seperator>*}+[)]
<y_seperator> ::= .|;
< Year> :: ={ 1900 |19001|19002|….|current_year} [a|b|…]

4.3 Document Parser

The document parser's process is to mine all necessary citation and reference data from the research article as input and import it into the database. We fitted the parser of the current prototype to the PMC OAS. Therefore, the

Table 1. List of citation extraction techniques

Algorithm/ Mechanism	Author, year	Reported performance	Comments on results
Integrated evidence-based algorithm	Powley and Dale, (2007)	The pattern list is done from a variety of conferences, workshops, and journals; it is still evident that the mixture of document styles in such a corpus is limited. In this task the aim is to increase the algorithm's effectiveness.	The algorithm's effectiveness to a small range of documents
Knowledge based system	Giuffrida et al. (2000)	Reporting 87% accuracy	Concentrated only on the author's name
Graph-partitioning problem	Chen Ding(1999)	This method performs better approach when compared to previous work by citation graph partitioning, to find research papers on the web.	This method helps citation link information to be explored.
CitePlag's detection algorithms	Norman Meuschke(2012)	Finds the similarity factors for the scoring of identify citation.	Used only the prototype model.

current version of CitePlag available for download and to process NXML-texts. In the future, and the prototypic parser with a component that uses the open-source citation extraction tool ParsCit[11].

5. Feature Extraction

The collected citations are used to dig out features for training sentiment classifiers. We used the existence of an n-gram as a binary feature, while for common information retrieval purposes, the frequency of a polarity existence in a more suitable feature, since the overall citation may not necessarily be indicated through the repeated use of polarity words. The list of feature extraction techniques are given in Table 2.

The *n*-gram is of use in features for feature mining. It is a contiguous chain of *n* objects from a specified chain of text. It could be any combination of letters. The *n*-grams typically are collected from a text, *n*-gram features gets sentiment cues in text. Fixed *n*-grams are exact sequences. Variable *n*-grams are extraction context capable of representing additional sophisticated linguistic phenomena. *N-gram* attributes can be classified into two types: 1. Fixed *n*-grams are sequences taking place at token level. 2. Variable *n*-grams are extraction context able of representing additional sophisticated linguistic phenomena.

The process of obtaining *n*–gram can be given as in the steps below:

- Filtering – removing URL Links (e.g. http://data.com) as reference.
- Tokenization – fragmenting sentence by dividing it by spaces and punctuation marks and forming bag of words.
- Removing Stop Words – eliminating ("a","an","the") in collected citation from the sack of words.
- Constructing *n*-grams – from consecutive words, we make a collection of n-grams by eliminating (such as "not") attached words which prefix or postfix it. For example, a statement "his works do not give right solution". Such a method improves the efficiency of the classification.

6. Sentiment Classification

Sentiment classification broadly refers to binary categorization, multi-class categorization, regression and ranking. It generally consists of two main processes, including senti ment polarity assignment and sentiment intensity assignment. Sentiment polarity assignment describes the analysis part, whether a context has a positive, negative or neutral semantic orientation. Sentiment intensity assignment explains another way of analysis, whether the positive or negative sentiments are soft or tough. There are many processes in order to reaching the objectives of sentiment analysis. These processes contain sentiment or opinion detection, polarity classification and unearthing of the opinion's result.

There are many tools and techniques for tackling the complexity in order to perform sentiment analysis. There are several methodologies used in order to achieve sentiment classification are;

Table 2. Lists of feature extraction techniques

Algorithm/Mechanism	Author, year	Reported Performance	Comments on results
Candidate Identification and Frequent Pattern Generation (CI-FPG)	Tanvir Ahmad (2012).	FP Growth Algorithm and rule based system is implemented to extracts all features,	This method does not filter the unwanted features
Feature Relation Network (FRN)	Ahmed Abbasi (2001)	proposed to efficiently permit the insertion of extended sets of mixed n-gram features	Developed hybrid technique to extract feature
Negation word detection	Michael Wiegand (2010)	Used syntactic knowledge for negation of expressions	Only negation modeling is considered
Multiple base classifiers and stacking with SVM	Zan Huang (2002)	Predefined features together with weighting mechanisms are used.	Vector method is implemented.
Correlation-based Feature Selection	Hall,M (1997)	feature selection algorithm (CFS) that operates autonomously of any induction algorithm	outcome is easier and understandable
Bi-Normal Separation' (BNS)	Forman,G (2004).	compared twelve feature selection techniques	Novel feature selection metric introduced

1. Classification with respect to term frequency, *n*-grams, negations , 2. Detection of the semantic orientation of words by lexicon, statistical techniques and training documents, 3. Identification of the semantic orientation of the context and phrases, 4. Identification semantic orientation of the documents, 5. Object feature extraction, 6. Comparative sentence identification.

6.1 Polarity Assignment

Sentiment polarity classification is a binary classification process where an opinionated context is labeled with a positive or negative sentiment. Sentiment polarity classification can also be termed as a binary decision process. The input given to the sentiment classifier can be opinionated or sometimes not. When a research article is given as an input for analyzing and classifying it as a positive or negative citation context is measured by text categorization task. Moreover, this piece of information can be positive or negative citation, but not necessarily subjective. Summarizing citations in order to gather information on to why the reviewers liked or disliked the research article is a new way of mining opinion. Types of citation polarity instance are given in Table 3.

6.2 Intensity Assignment

While sentiment polarity assignment deals with analyzing, whether a context has a positive, negative or neutral semantic orientation, sentiment intensity assignment deals with analyzing, whether the positive or negative sentiments are soft or strong. Consider the two phrases "It doesn't produce required accuracy" and "Its accuracy is bad", where, these sentences would be assigned a negative semantic orientation but the latter would be considered more intense than the first. Effectively classifying sentiment polarities and intensities entails the use of classification methods applied to linguistic features. There are different classifications methods have been employed for opinion mining; Support Vector Machine (SVM) has outperformed when compare to different techniques including Naive Bayes, Decision Trees.

6.3 Machine Learning Approaches

Machine learning provides a solution to the classification problem and involves two steps:

1. Learning the model from a corpus of training data,
2. Classifying the unseen data based on the trained model.

Examples or text classification techniques are given in Table 4.

6.3.1 Naive Bayes Classificatio

It is an approach to text classification that the probabilistic learning technique that assumes terms appens autonomously. Given a collection of N citation patterns $\{P_j\}N_{j=1,}$ where each pattern is stands for as a sequence of T terms $Pj = \{t_1, t_2, \ldots. t_T\}$, the probability of a pattern Pj arises in class C_k is specified as:

$$P(C_k|p_j) = P(C_k) \prod T_{i=1} P(t_i|C_k)$$

Where $P(C_k|p_j)$ is the conditional probability of term t_i arising in a Pattern of class C_k and $P(C_k)$ is the prior

Table 3. Types of Citation Polarity

Citation Polarity	Citation Sentences
Positive	The opinions expressed by users are an important factor taken into consideration by product vendors vendors (Hoffman 2008)
Negative	Existing reported work[12] falls short of this high accuracy
Neutral	[12] report on heuristics for digging out citations from ACM papers, reporting precision of 0.53 , based on arbitrarily selected papers

Table 4. List of sentiment classification

Algorithm/Mechanism	Author, year	Reported Performance	Comments on results
Cross-domain sentiment classification	Bollegala,D (2012)	Used integrate document level sentiment labels in the context vectors	Used single- and multisource domain.
Rule-based, SVM and Maximum Entropy	Naradhipa,A.R (2012)	Processed several steps like text preprocessing, feature extraction, and classification	Shown that SVM generate 83.5% accuracy
Imbalanced class distribution and uncertainty method introduced	Shoushan Li (2012)	Used disjoint feature subspaces and samples for manual annotation	Reduces the human annotation

probability of a pattern arising in class C_k. P ($t_i|C_k$) and P (C_k) are calculated from the training data.

6.3.2 Decision Trees

Zhang Wei (2010) Decision trees are planned with the use of a hierarchical partition of the underlying data space with the use of a several text features. The hierarchical partition of the data space is planned for development of class partitions which are more twisted in terms of their class sharing. For a given text sentence, we find out the partition that it is mainly likely to belong to and use it for the purpose of classification.

6.3.3 Support Vector Machines

A Support Vector Machine is a supervised learning technique with different attractive features that make it an accepted algorithm. It has a solid theoretical base and executes classification more accurately than the majority of other algorithms in many applications. Several researchers have stated that SVM is perhaps the most accurate method for text classification (Liu, 2011).

The basic preparation behind the training method is to get a highest margin hyper plane, represented by vector W, which not only breaks up the pattern vectors in one class from those in the other, but for which the partitions or margin, is as big as possible. This corresponds to a constrained optimization problem; letting C_j belongs to {1,-1} be the correct class of pattern P_j, the output can be specified as below equation:

$$W = \sum_j \alpha_j c_j P_j, \ \alpha_{j>=0}$$

Where, the α_j's (Lagrangian Multipliers) are performed by solving a double optimization problem. Those P_j such that α_j larger than zero are called support vectors, since they are the only pattern vectors contributing to W.

Classification is test example consists simply of determining which side of W's hyper plane they fall on.

7. Analysis of Trends in Sentiment Analysis

The datasets used for performing citation analysis have been taken from the Google scholar database. Since, Google scholar has a larger number of high impact journals, those papers have been referred to in different papers. In addition, the majority of the referencing

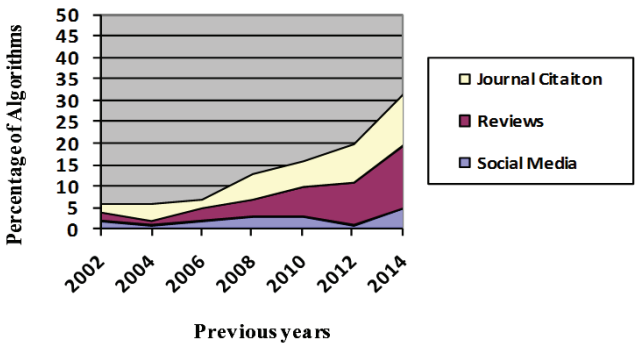

Figure 3. Percentage of algorithms targeting several areas over the previous years.

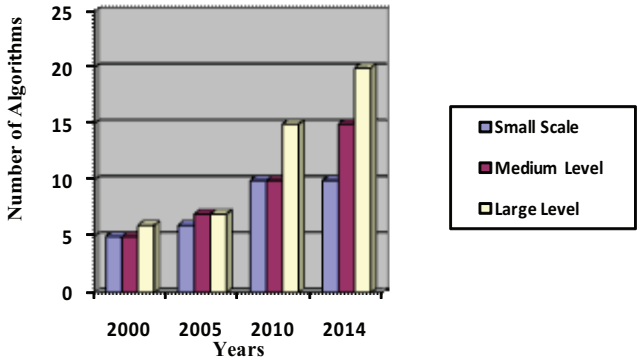

Figure 4. Number of algorithms with several scalability levels over the previous years.

papers are straightforwardly accessible with their full text format. We have taken 25 research papers in the domain of sentiment analysis on scientific researcher reviews. We have also collected the referencing papers of these 10 citations. In this section, we perceive the emerging trends; compare the different methods that have been proposed for sentiment analysis. We now discuss some trends that emerge when analyzing the recent publications on sentiment and compare them with the list of papers with different technique. We assign the papers to different classes under different dimensions: based on the employed algorithms, datasets used for testing and target areas. In Figure 3 and Figure 4 shows the scalability levels and several of the properties of the papers we used for the above analysis.

8. Conclusion and Future Work

This survey discusses various approaches in sentiment analysis. It provides a detailed overview of different

approaches in citation analysis and article ranking with potential challenges of sentiment mining that makes it a difficult task. We studied the citation analysis for journal citation sentiment mining. Several citation extractions, preprocessing technique and some of the machine learning techniques like Naïve Bayes, Decision Tree and Support Vector Machines have been discussed. These have emerged as important areas of sentiment mining. Our survey expose these trends in citation analysis, the recent developments in sentiment analysis its related subtasks are also presented. The state of the art of existing approaches has been integrated to extract the citation. We improve the additional features when finding the polarity between the citations of our future work.

9. References

1. Powley B, Dale R. High accuracy citation extraction and named entity recognition for a heterogeneous corpus of academic papers. International Conference on Natural Language Processing and Knowledge Engineering: NLP-KE'07; Beijing. 2007 Aug 30-Sep 1. p. 119–24.
2. Ding C, Chi CH, Deng J, Dong CL. Citation retrieval in digital libraries. IEEE International Conference on Systems Man and Cybernetics, IEEE SMC'99 Conference Proceedings; Tokyo. 1999; 2: p.105–9.
3. Huang Z, Chung W, Ong TH, Chen H. A graph-based recommender system for digital library, Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL'02; 2002. p. 65–73.
4. Dong C, Schafer U. Ensemble-style self-training on citation classification. Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand: Asian Federation of Natural Language Processing; 2011. p. 623–31.
5. Koppel M, Schler J. The importance of neutral examples for learning sentiment. Computational Intelligence. 2006; 22(2):100–9.
6. Zhang W, Li P, Zhu Q. Sentiment classification based on syntax tree pruning and tree kernel. 2010 7th Web Information Systems and Applications Conference IEEE; Hohhot. 2010 Aug 20-22. p. 101–5.
7. Zhou G, Li X, Li S, Ju S. Active learning for imbalanced sentiment classification. Proceedings of the International Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning; 2012. p. 139–48.
8. Parthasarathy G, Tomar DC. Trends in citation analysis. ICCD, Advances in Intelligent Systems and Computing. India, Springer. 2015; 308; p. 813–21.
9. Parthasarathy G, Tomar DC. Optimized Prune Based Data Mining (OPBDM) for distributed databases: An adaptive approach. Journal of Theoretical and Applied Information Technology. 2014 Sep; 67(3):717–24.
10. Chandrakala S, Sindhu C. Opinion mining and sentiment classification: A survey. ICTACT Journal on Soft Computing. 2012 Oct; 3(1):420–7.
11. Councill IG, Giles CL, Kan MY. ParsCit: An open-source CRF reference string parsing package. Proceedings of LREC, European Language Resources Association (ELRA). 2008. p. 661–7. Available from: http://aye.comp.nus.edu.sg/parsCit/
12. Bergmark D. Automatic extraction of reference linking information from online documents. Cornell Digital Library Research Group. 2000.
13. Kang M, Shin JD, Kim B. Automatic subject classification of Korean Journals based on KSCD. Indian Journal of Science and Technology. 2015 Jan; 8(S1):452–6.