iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4282

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

A Survey on Automatic Semantic Subject Indexing of Documents using Big Data Analytics

K.Swanthana[1], K.Swapnika[2]
[1, 2]*Gokaraju Rangaraju Institute of Engineering and Technology, Telangana, India*

Abstract: ***The automatic subject indexing of documents is prevailing issue due to the increase in quantity and diversity of digital documents available to end users. So there is a need for effective and efficient indexing and retrieval techniques. Indexing is a crucial aspect that allows the documents to be located quickly. Instead of full-text indexing on documents, the metadata such as title of publication and abstract may be considered for performance and accuracy. To retrieve the documents which are contextually related by annotating the massive collection with only the title and abstract, whereas individual words provide unreliable evidence about the conceptual topic or meaning of a document. Hence, the available approaches cannot meet several challenges of data in terms of processing. This results in inefficient query results. There is a need for the design of indexing strategies that can support. There are various indexing strategies which are utilized for solving Big Data management issues, and can also serve as a base for the design of more efficient indexing strategies. The aim is to explore methods of indexing and retrieving the documents based on the different query search types, by utilizing some of the subject indexing strategy for Big Data manageability by identifying some of the challenges of existing strategies. The existing strategies like, Vector Space Models, Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Logistic Regression, Linear Discriminant Analysis, Naïve Bayes and Logistic Regression which have their own challenges. This paper will describe about some of the Automatic subject-indexing approach's applied to retrieve subject specific Document and presents the characteristics and challenges involved.***
Keywords: Big Data; subject indexing; Documents; Context; Query; Retrieval;

I. INTRODUCTION

During the past years, there has been an extraordinary growth in the amount of digital information available to common end-users. Factors contributed to this growth include the world-wide proliferation of the Internet, the economical cost of digitizing information contents into computerized forms, and a general increase in computer literacy and accessibility to a large and growing number of people around the world.

Automatic Semantic indexing occupies an important position in document classification and information retrieval. Document ranking largely depends on measuring the semantic similarity of query-document pairs. There are many professional terms (multi-labels) in the documents, but considering only the title and abstract information of the document and the high correlation between different labels/terms is observed. The problem is instead of the full-text indexing on documents using metadata such as title of publication or paper for subject indexing for increasing the performance. Concept search as an alternative to keyword search which is an important means in information retrieval. It assumes that the terms that appear frequently in the same document are likely to be related to each other through unidentified concepts. By considering the hidden relationship of terms, the concept search tries to overcome the difficulties of synonymy and polysemy that emerge in the keyword search.

Semantic annotations are crucial for users of digital libraries as they enhance the search of scientific documents. Given the large amount of new publications, automatic annotation systems are a useful tool for human expert annotators working at digital libraries to classify the publications into categories from a (hierarchical) thesaurus. However, providing automated recommendations for subject indexing in such systems is a challenging task. This is partly due to the data from which recommendations may be generated. Often neither the full-text of a publication nor its abstract may be available. For instance, the digital library EconBiz contains only for 15% of the documents an abstract. Even when the content can be legally provided by the library to the end users, copyright laws or regulations of the publishers may prevent text mining. By collecting and processing PDFs for some Open Access documents, adds high computational requirements to the library. This puts annotation methods on demand that are based on data with better availability, such as the title. Previous work by Galke et al. [1], however, has shown that title-based methods considerably fall behind full-text methods considerably fall behind full-text methods in terms of performance when the number of samples for training is equal. If our classifier was a human expert, this would not be a surprising result.

©IJRASET (UGC Approved Journal): All Rights are Reserved

A full text contains more information and therefore also more indication of the publication's topic. A human expert will always make better annotations based on the full-text. In fact, the annotations that are used as gold-standard for automated subject indexing experiments are often created based on the full-text. However, we argue that machine learning algorithms work differently than a human. In contrast to a human, they often require hundreds of thousands or even millions of training data to yield satisfactory models [2]. These amounts of data are not always available in the real world.

One common reason is that human expertise is required for creating a large enough gold standard, which is expensive. For semantic subject indexing, the availability issues mentioned above come into play at prediction time, when a machine learning model is used in a productive system, and also during training. In effect, methods based on the full-texts have drastically less training data available than methods based on titles. This raises the question if title-based methods can potentially narrow the performance gap to full-text methods by fully incorporating all training data available.

II. RELATED WORK

In this section, we review previous literature relevant to our study. First, we discuss on title and abstract indexing. Finally, we discuss current methods for document classification.

Document classification denotes assigning an unknown document to one of predefined classes. This is a straightforward concept from supervised pattern recognition or machine learning. It implies 1) the existence of a labeled training data set, 2) a way to represent the documents, and 3) a statistical classifier trained using the chosen representation of the training set. Some classifiers are very sensitive to the representation, for example, failing to generalize to unseen data (overfitting) if the representation contains irrelevant information [4]. It would thus be advantageous to be able to extract only information pertinent to classification. However, some classifiers, such as Support Vector Machines [5], tolerate better irrelevant information. Either case, in general, it is computationally cheaper to operate the classifier in low dimensional spaces. In this paper we discuss document representation methods and approaches to reduce their dimensionality.

Automatic subject indexing performing linguistic analysis for matching document words expressed as terms in a controlled vocabulary (semantic tagging) and determining which of the matched vocabulary terms will best describe the document (topic ranking).

Semantic tagging: It is matching of words to meanings and a part of linguistic analysis. Linguistic analysis for the purpose of annotation consists of five steps: morphological analysis, part-of-speech tagging, chunking, dependency structure analysis and

Semantic tagging [6] in languages such as English, Spanish and French, a simplified form of semantic tagging can be performed by using a rule based stemming algorithm to normalize both document words and vocabulary terms [7]. This allows plural terms in the vocabulary. Inflected languages such as Finnish, Turkish, Arabic and Hungarian typically express meanings through morphological affixation. In highly inflected languages plural and possessive relations, grammatical cases, and verb tenses and aspects, which in English would be expressed with syntactic structures, are characteristically represented with case endings [8,9]. Compound words are also typical in inflected languages. Rule-based stemming does not work particularly well for semantic tagging: as an example, a semantic tagger for the Finnish language developed in the Benedict project used a sophisticated morphological analysis and lemmatisation tool as well as rules for handling compound words in order to attain high precision [9,10].

In topic ranking, machine learning methods have surpassed rule-based methods for determining the important topics of a document [11]. The TF×IDF method provides a baseline [12], which Maui [7] and its predecessors KEA [14] and KEA++ [13] have improved on by additionally using various heuristics. These tools can also perform topic indexing without the support of a controlled vocabulary, known as key phrase extraction. The previous Maui tests on English, French and Spanish documents have used a stemming algorithm for basic semantic tagging. In those languages, Maui has been found to assign subjects of comparable quality of those of human indexers [7]. A KEA-like approach for key phrase extraction of Arabic documents has also been found to perform well when part-of-speech analysis was incorporated into the candidate selection phase [16]. Other subject indexing tools for inflected languages include the Poka information extraction tool for Finnish [17], which has been used, e.g., in the Opas system to assign concepts from the Finnish General Upper Ontology to question answer pairs [18]. It is used by many Finnish news websites for automatically generating links to related content. However, neither tool has been evaluated in academic literature.

III. METHODS

A. *Vector Space Models*

In vector-space model, a document is theoretically represented by a vector of index terms exported from the text, with related weights which represent the importance of the index terms in the document and within the whole document collection; likewise, a

©IJRASET (UGC Approved Journal): All Rights are Reserved

query is modeled as a list of index terms with related weights that represent the importance of the index terms in the query. Vector space model has four main methods that will be the core of our study to find the best method among them. These four techniques are Inner Product, Cosine similarity, Dice Similarity, Jaccard Similarity.

It is used for classification, it map's the frequency of co-occurrences of words in the documents into a vector space and utilize similarity measures to match search terms with a document. They tend to over-fit a corpus set and suffer from the difficulties of synonymy and polysemy which are in keyword searches.

B. Latent Semantic Indexing

Latent semantic indexing (LSI), sometimes also called latent semantic analysis (LSA), is an indexing and retrieval method that is based on the principle that words used in the same context tend to have the same meaning. LSI uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. This ability to extract the meaning by establishing associations between terms of text is a key feature of LSI [20]. One benefit of LSI is that it solves two of the most problematic constraints in literal keyword searches: different words have the same meaning (synonymy) and the same word having different meanings (polysemy). These issues lead to mismatches in vocabulary used between the documents in the collection and in the queries being performed, resulting in both low precision and recall.

Another benefit of LSI is that, since it is a pure mathematical method, it does not rely on any knowledge about the text. This means that LSI works well with any language and it can even be used to find documents across languages, especially in scientific collections where a lot of the terminology is the same between languages. LSI is also very tolerant to noise, like misspelled words, and it adapts well to changes in the terminology used in the data collection.

It uses SVD technique and projects the high-dimensional data into a lower dimensional space to overcome the over-fitting problem. LSA compared to VSMs, provides success in alleviating the difficulties of the synonymy, but it still experiences difficulties in output interpretation which are obtained from the analysis.

C. Probabilistic Latent Semantic Analysis (PLSA):

Probabilistic Latent Semantic Indexing is a novel approach to automated document indexing which is based on a statistical latent class model for factor analysis of count data. Fitted from a training corpus of text documents by a generalization of the Expectation Maximization algorithm, the utilized model is able to deal with domain specific synonymy as well as with polysemous words. In contrast to standard Latent Semantic Indexing (LSI) by Singular Value Decomposition, the probabilistic variant has a solid statistical foundation and defines a proper generative data model.

It is a generative model based on LSA; it provides clarity in the interpretation of the output values as they have meanings of probability. PLSA also deals with difficulties of parameter interpretation as it assigns a single probability measure to a document with respect to a topic variable and it over-fits training datasets because the number of parameters grows larger as number of documents increases in a corpus.

D. Logistic Regression (LR)

It is mainly used in cases where the output is Boolean (true/false). It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable (coded 0, 1).

There are two main Advantages, first is you can include more than one explanatory variable (dependent variable) and those can either be dichotomous, ordinal, or continuous. The second is that logistic regression provides a quantified value for the strength of the association adjusting for other variables (removes confounding effects). The exponential of coefficients correspond to odd ratios for the given factor.

It is a classification algorithm limited to only two-class or binary classification problems. If there are more than two classes then Linear Discriminant Analysis is preferred, Linear classification technique used. It becomes unstable when the classes are well separated and when there are few examples from which to estimate the parameters. Logistic regression is a classification algorithm traditionally limited to only two-class classification problems. If we have more than two classes then Linear Discriminant Analysis is the preferred linear classification technique.

©IJRASET (UGC Approved Journal): All Rights are Reserved

E. Linear Discriminant Analysis(LDA):

There are many approaches for obtaining topics from a text such as – Term Frequency and Inverse Document Frequency. Latent Dirichlet Allocation is the most popular topic modeling technique and in this paper, we will discuss the same.

LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution. Given a dataset of documents, LDA backtracks and tries to figure out what topics would create those documents in the first place.

This algorithm is used for classification predictive modeling problems. It addresses all the limitations of PLAS and LR methods. It has become one of the popular probabilistic text modeling technique in machine learning. LDA models a document as a mixture of multiple topics. The advantage is it encourages results with topics modeling.

F. Naïve Bayes (NB)

Naive Bayes classifiers are linear classifiers that are known for being simple yet very efficient. The probabilistic model of naive Bayes classifiers is based on Bayes' theorem, and the adjective naive comes from the assumption that the features in a dataset are mutually independent. In practice, the independence assumption is often violated, but naive Bayes classifiers still tend to perform very well under this unrealistic assumption. Especially for small sample sizes, naive Bayes classifiers can outperform the more powerful alternatives. Being relatively robust, easy to implement, fast, and accurate, naive Bayes classifiers are used in many different fields.

In the Document system, the stem forms of words occurring in the training documents were used as the features to represent each document. The basic steps in the naive Bayes method are: Training: •Identify the individual stem words occurring in all the training documents in the training set. • Generate the feature vector for each document in the training document set and store it along with the correct indexes in the knowledge base. • Calculate the probability for each index. It is a generative model. Naïve Bayes classifier is a probabilistic linear classifier uses Bayes Theorem and has strong independence among features. Here we compute the probability the document d being in a class c, finding the "best class" is the main goal.

IV. COMPARISON OF INDEXING STRATEGIES

Indexing Strategies	Data-type	Query-type
Vector space model	Documents	Keyword search
LSI	Multimedia data, spatial data (textual data)	Keyword search
PLSI	Textual data	Keyword search
Latent Dirichlet Allocation (LDA)	Documents	Concept Search
Naïve Bayes (NB)	Documents, text data	Concept Search
Logistic Regression (LR)	Documents, text data	Concept Search

Table I: Indexing Strategies and Query-Types

©IJRASET (UGC Approved Journal): All Rights are Reserved

Indexing	Strategies	Properties Challenges
Vector Space Model-Classification	VSM- it represents both documents and queries with high-dimensional vector - measures the cosine of the angle between the two vectors in the vector space for classification.	-There is no real theoretical basis for the assumption of a term space. -it is more visualized. - Terms are not independent of all other terms.
LSI	LSI - Uses data and meaning of data for indexing - Presents accurate query results (since it uses more information)	-Demands high computational performance - Consumes more memory Space
PLSI	PLSI-it provides clarity in the interpretation of the output values - it assigns a single probability measure to a document with respect to a topic variable	- over-fits training datasets because the number of parameters grows larger as number of documents increases in a corpus.
Latent Dirichlet Allocation (LDA)	-LDA is a matrix factorization technique. -LDA generates words based on their probability distribution. -it encourages results with topics modeling.	-One challenging issue of LDA is to select the optimal number of topics in LDA model. -No such topic selection method which considers the density of each topic and computes the most unstable topic structure.
Naïve Bayes (NB)	- It is a probabilistic linear classifier uses Bayes Theorem -It is strong feature independent. - Naive Bayes is a good tool which is fast, robust and relatively insensitive to missing values and even data imbalance problems.	-No attribute or feature independence. - When attribute is continuous, computing the probabilities and frequency counts is not possible. - Incomplete training data: when class conditional probability is zero, the whole construction collapses.
Logistic Regression	-	- Limited to only two-class or binary classification problems. - Logistic regression combines both binomial and normal distribution, this can sometimes cause problems.

Table II: Characteristics of Indexing Strategies

V. CONCLUSION AND FUTURE SCOPE

The existing techniques have their own challenges. Hence, we want to implement a effective subject document indexing technique. The objectives in future are to implement an effective contextual subject indexing on documents, to lower the data processing time and to enhance the Performance of the indexing and retrieval using big data analytics.

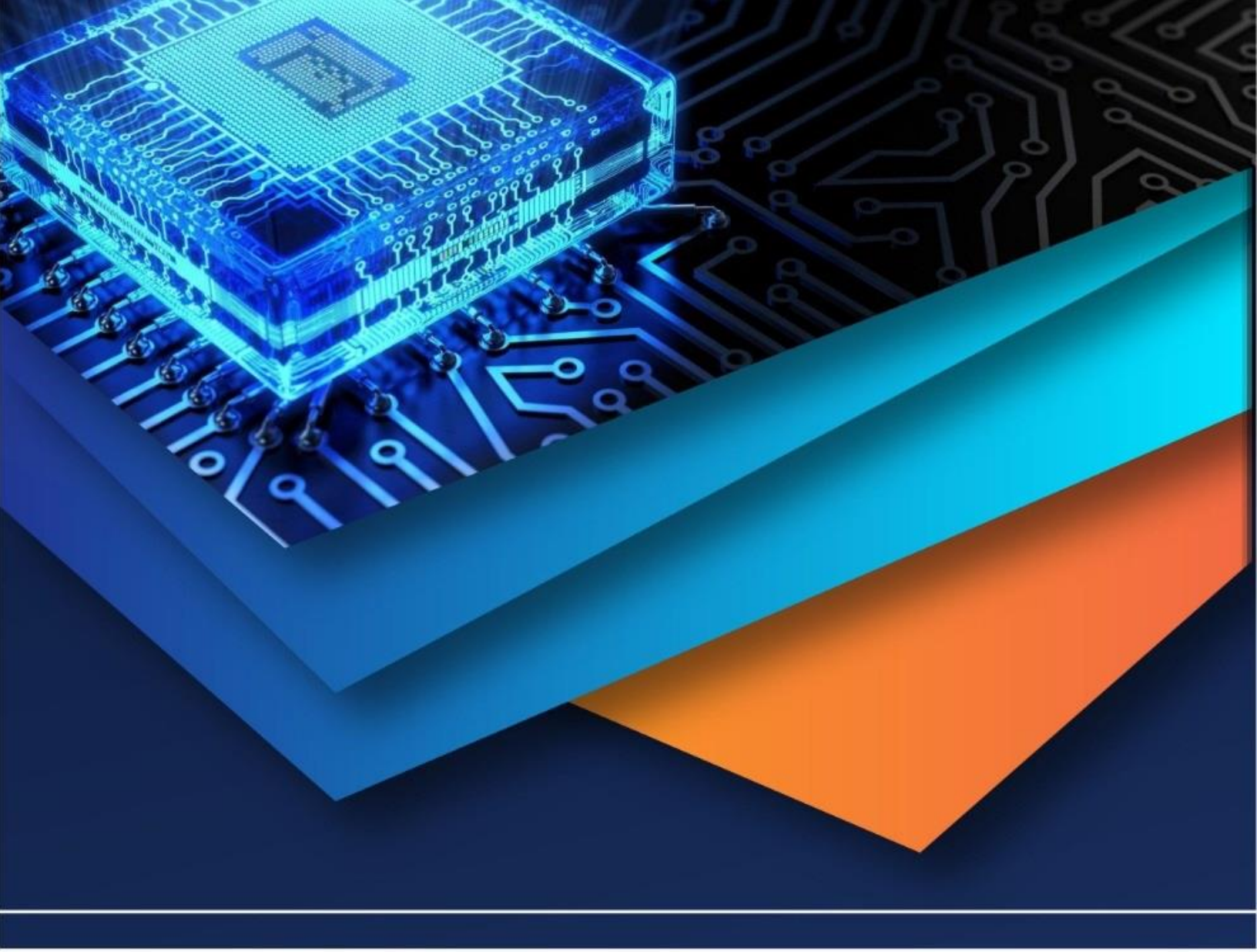
REFERENCES

[1] Lukas Galke, Florian Mai, Alan Schelten, Dennis Brunsch, and Ansgar Scherp. 2017. Using Titles vs. Full-text as Source for Automated Semantic Document Annotation. In K-CAP. 9.
[2] Damien Brain and G Webb. 1999. On the effect of data set size on bias and variance in classification learning. In Australian Knowledge Acquisition Workshop, AI'99. 117–128.
[3] A. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum"A survey on indexing techniques for big data: taxonomy and performance evaluation." Knowledge and Information Systems 46.2 (2016): 241-284.
[4] Christopher M. Bishop. Neural Networks for Pattern Recognition. Oxford University Press,Oxford, New York, 1995.
[5] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Claire N´edellec and C´eline Rouveirol, editors, Proceedings of CML-98, 10th European Conference on Machine Learning, pages 137–142, Chemnitz, DE, 998. Springer Verlag, Heidelberg, DE.

©IJRASET (UGC Approved Journal): All Rights are Reserved

[6] Buitelaar, P., Declerck, T.: Linguistic Annotation for the Semantic Web. In: Annotation for the Semantic Web, pp. 93–110. IOS Press, Amsterdam (2003)
[7] Medelyan, O.: Human-competitive automatic topic indexing. Ph.D. thesis, University of Waikato, Department of Computer Science (2009)
[8] Oflazer, K., Kuruöz, I.: Tagging and Morphological Disambiguation of Turkish Text. In: Proceedings of the Fourth Conference on Applied Natural Language Processing(1994)
[9] Löfberg, L., Archer, D., Piao, S., Rayson, P., Mcenery, T., Varantola, K., pekk Juntunen, J.: Porting an English semantic tagger to the Finnish language. In: Proceedings of the Corpus Linguistics 2003 Conference (2003)
[10] Löfberg, L., Piao, S., Nykanen, A., Varantola, K., Rayson, P., Juntunen, J.P.: A semantic tagger for the Finnish language. In: Proceedings of Corpus Linguistics 2005 (2005)
[11] Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM ComputingSurveys 34(1), 1–47 (2002)
[12] omorfi Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)
[13] Medelyan, O., Witten, I.H.: Thesaurus Based Automatic Keyphrase Indexing. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (2006)
[14] Witten, I.H., Paynter, G., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: Practical Automatic Keyphrase Extraction. In: Proceedings of Digital Libraries 1999 (1999)
[15] Pala, N., Çiçekli, I.: Turkish Keyphrase Extraction Using KEA. In: Proceedings of the 22nd International Symposium on Computer and Information Sciences, ISCIS 2007 (2007)
[16] El-Shishtawy, T., Al-Sammak, A.: Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools (2009)
[17] Valkeapää, O., Alm, O., Hyvönen, E.: Efficient content creation on the semantic web using metadata schemas with domain ontology services (System description). In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp.819–828. Springer, Heidelberg (2007)
[18] Vehviläinen, A., Hyvönen, E., Alm, O.: A semi-automatic semantic annotation and authoring tool for a library help desk service. In: Emerging Technologies for Semantic Work Environments: Techniques, Methods, and Applications, pp. 100–114. IGI Group, Hershey (2008)
[19] Pennanen, P., Alatalo, T.: Leiki – a platform for personalized content targeting. In: Proceedings of the 12th ACM Conference on Hypertext and Hypermedia, HYPERTEXT 2001 (2001)
[20] S. Deerwester, S. T. Dumais, T. K. Landauer. "Indexing by latent semantic analysis". Journal of the American Society of Information Science. Vol. 41. 1990. pp.391- 407.

©IJRASET (UGC Approved Journal): All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)