2012 10th International Conference on Frontiers of Information Technology

KeaKAT – An Online Automatic Keyphrase Assignment Tool

Rabia Irfan, Sharifullah Khan, Irfan Ali Khan, Muhammad Asif Ali
School of Electrical Engineering & Computer Science (SEECS),
National University of Sciences and Technology (NUST), Islamabad, Pakistan
{09msitrirfan, sharifullah.khan, 08bitikhan ,08bitasifa}@seecs.edu.pk

***Abstract*—Kea++ is a well known tool for assigning keyphrases to documents. But Kea++ contains noise and irrelevant terms in the keyphrase result set. The extended refinement methodology was developed to fine tune the results of Kea++ for multiple domains. However using Kea++ and its refinement as a system for assigning keyphrases to documents is not simple for users of a domain other than computing. The system needs to be installed and configured. It does not have any GUI to facilitate users in assigning keyphrases. The objective of the KeaKAT is to develop a web-based keyphrase assignment tool to facilitate users in assigning relevant keyphrases to their documents online. KeaKAT saves users not only from installing and configuring the system, but also improves the usability of the system.**

***Index Terms*—Keyphrase assignment, information extraction, usability, classification system**

I. Introduction

Keyphrases are the words that present the concise summary of a document or text [1], [5], [11]. They can be used in variety of applications that involve organization and management of the huge amount of information. They can be helpful in browsing document collections [8], can be used as metadata [16], can be used to index document collections [16], [7] and can assist in classification and clustering of document collections [10]. Because of their usage in different applications, many tools were develop that can be helpful in automatically generating keyphrases. Two main approaches were used; one is extraction of keyphrase from a document text known as keyphrase extraction and other is alignment of document with a classification system/taxonomy known as keyphrase assignment. Kea++ [14] is a well known tool that can perform both keyphrase assignment and extraction based on a given input. However the output produced by Kea++ contains noise and irrelevant terms. The work done by [4] proposed refinement methodology that takes Kea++ assigned keyphrases as input and generates refined keyphrases. The methodology exploits the hierarchical structure of a taxonomy [15], [6] as well as common heuristics to fine tune the result of Kea++. The refinement methodology was extended in the work [9]. The extended refinement methodology aimed to improve and generalize the refinement methodology for multiple domains.

Both Kea++ and the extended refinement methodology work together to produce better results for keyphrase assignment to documents. Kea++ and its refinements as a system need installation and configuration for assigning keyphrases to documents accurately. There is no graphical user interface (GUI) that has been provided to facilitate users in using the system. Automatic keyphrase assignment techniques are not only helpful for computing experts, but can be equally important and applied in other disciplines of academia, particularly in library sciences. The users belonging to fields other than computing are not computer experts. Most of the time they are not comfortable in using techniques that involve too much of the computer understanding, neither they bother themselves to understand the technical details of such systems. Therefore the available automatic tools are not widely used by academia. The objective of the KeaKAT is to improve the usability of the automatic keyphrase tools. KeaKAT is a web-based automatic keyphrase assignment tool. The tool uses Kea++ and the extended refinement methodology in background for keyphrase assignment. KeaKAT facilitates users to train and test Kea++ and also applies refinement in the background for assigning relevant keyphrases to their documents online. It saves users from installing and configuring the system and facilitates users through GUI to get their job easily done. It improves the usability of the system.

The rest of the paper is organized as follows: Section 2 discusses the related work and existing systems. Section 3 explains the architecture, working of the proposed tool. Comparative analysis of the existing systems with the proposed system is described in Section 3. Section 4 concludes the paper and discusses the future work.

II. Related Work

This section discusses existing systems for automatic keyphrase assignment. Kea [17] and its later version Kea++ [14] are famous tools developed at the University of Waikato for performing the task of keyphrase generation automatically. Kea is a machine learning based tool and it works in two phases; training phase and extraction phase. Initially Kea was used to extract keyphrase from documents later on it was extended to perform keyphrase assignment and known as Kea++. Kea++ also works in two phases like Kea i.e. training and extraction. During each phase it works in two sub steps: candidate identification and filtering. During candidate identification step language dependent techniques such as: input cleaning, stemming etc are applied to form pseudo phrases. During the filtering step, those keyphrases are identified which are the most suitable candidates based on four features: Term

978-0-7695-4927-9/12 $26.00 © 2012 IEEE
DOI 10.1109/FIT.2012.14

frequency–inverse document frequency (tf×idf), phrase's first occurrence, length of a candidate phrase in words and node degree. Then by applying Naive Bayes algorithm a training model is generated. Extraction phase uses the model to assign keyphrases to the document.

Keyphrase assignment task can also be performed by Maui [13] along with keyphrase extraction and tagging. Similar steps like that of Kea++ were followed to assign keyphrases to the document from taxonomy. However filtering stage utilizes additional features along with those used by Kea++. Like Kea++ it first builds a model based on training documents and based on this model, applies extraction and filtering on test documents to produce related keyphrases. Through Maui, one can perform the desired task of assignment, extraction or tagging based on the given input through one single algorithm. Maui has been successfully tested in computer science, agricultural and medicine domains, as well as on blog posts and news articles. However the real contribution of Maui is its assignment of terms from Wikipedia in the absence of domain specific controlled vocabulary. Maui is also available as web based system[1].

Agrotagger is a keyphrases assignment tool for agriculture, livestock, forestry and fisheries that uses Kea++ engine along with the Agrovoc thesaurus. It is a pluggable module developed with collaboration of Food and Agriculture Organization of the United Nations (FAO). It could be used as an add-on to leading repositories such as DSpace and advanced management systems like Drupal and Joomla to automatically tag documents[2]. Agrotagger uses Agrotags as candidate keywords for documents. Agrotags are a proper subset of Agrovoc. Concepts are selected from Agrovoc as Agrotags based on their utility in a tagging scheme as well as their popularity. Agrotagger works in three steps. It first identifies the occurrence of Agrovoc terms in the document, and then replaces them with an equivalent Agrotags term. In the end, it chooses the candidate keywords from among them based on the application of the statistical techniques. Agrotagger is also available as a web service[3] to automatically assign keyphrases for agricultural documents.

III. Proposed System

A. Architecture

We have adopted layered approach for the proposed system and kept the data layer separate from the business logic and application layers as shown in Figure-1.

1) Application tier: The top most level of the system is the application tier. The main function of this tier is to provide user interface. A user accesses various features of KeaKAT through application tier.

In order to train the system, a user can upload a training dataset of his own choice. The dataset is passed to the data tier and the business logic tier for further processing. Any interaction done by a user with KeaKAT is responded through the web server in this layer. A user can upload a document for which keyphrases are needed to be assigned. He can download the result set in XML format through this layer. A user is also allowed to enter his feedback for the final selection of keyphrases of his own choice from taxonomy.

[1]http://maui-indexer.appspot.com/; Last viewed: 29 June, 2012
[2]http://agropedia.iitk.ac.in/?q=content/agrotagger-version-ii
[3]http://agropedialabs.iitk.ac.in/Tagger/ ; Last viewed: 29 June, 2012

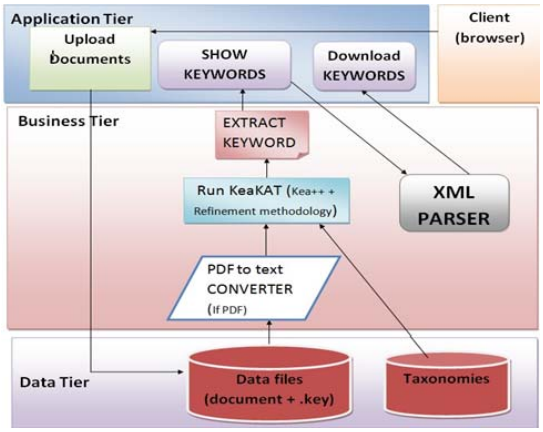

Figure 1. Architecture of KeaKAT

2) Business logic tier: All logical decisions of the system occur in the business tier. An input document format and taxonomy format is checked here. The allowed document formats are PDF and text. When a user uploads documents and taxonomy, then they are checked for their correct format. If they do not have the correct format, then the user is notified and is asked to upload the allowed type of documents. All PDF files are converted to text. This is done through a *PDF to text* converter. After all files are converted to text format and appropriate taxonomy has been selected, then KeaKAT is run. It generates keyphrases for test documents. A user can directly provide feedback from taxonomies using an XML parser. We parse the taxonomies to HTML and then pass it to the application layer in order to display it in a tree structure. An XML file of the result set is generated in this layer.

3) Data Tier: The data tier contains all folders and the database which are used in the system. The storage of training or testing datasets provided by a user is handled in this layer. The taxonomy used in the system is stored and catered in this layer.

B. Working

1) Start KeaKAT and Log-in: A user can access KeaKAT online by typing its URL in the address bar of a web browser. To use this application, a user must first sign up. The user provides an email address. Relevant log-in information are then sent to the given email address automatically. After successfully logging into the system, the user is directed to the home page. Guidelines for users are available here to increase the user learnability of the system.

2) KeaKAT Training: Since Kea++ is machine-base learning system, it needs training before testing. There are two steps i.e. Step-1 and Step-2 in KeaKAT for training purpose.

a) Step 1 : In Step-1, a user uploads training documents along with their manually assigned keyphrases. A user can

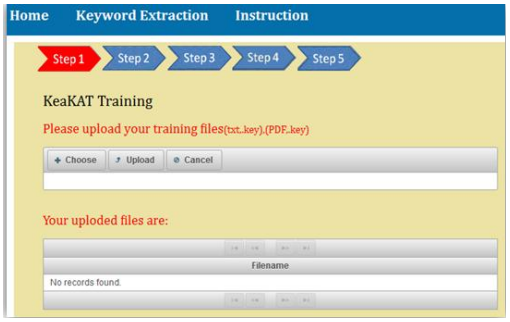

Figure 2. KeaKAT Training

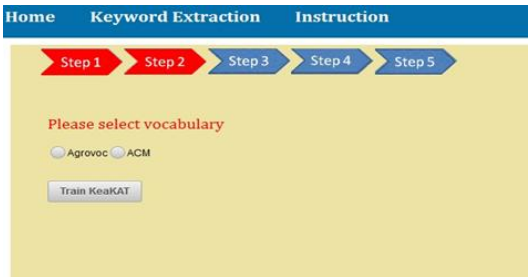

Figure 3. Vocabulary Selection

upload one document at a time for training or a bulk of documents. KeaKAT uses these documents and their keyphrases in training as shown in Figure-2.

b) Step 2 : During this step, a user selects a vocabulary/taxonomy. Currently there are two vocabularies available in our system: Agrovoc and ACM. After vocabulary selection, a user trains KeaKAT by pressing the "Train KeaKAT" button as shown in the Figure-3.

3) Keyphrase Assignment: Once Kea++ is trained, then it can assign relevant keyphrases to a document.

a) Step 3 : In this step, a user can upload a single document or more than one document for assigning keyphrases to them. Moreover a user can also have the option of copying and pasting the text in the text area besides uploading documents as shown in Figure-4. A user can upload only four documents at a time currently in our system for keyphrase assignment. We have imposed this limit to avoid overloading of the system.

4) User Feedback :

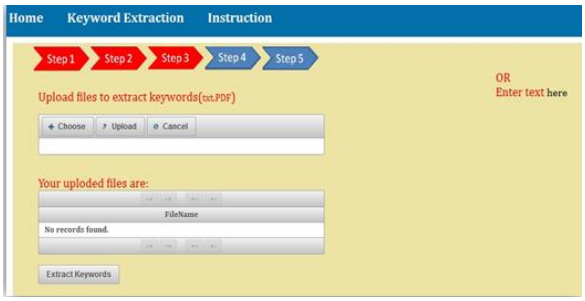

Figure 4. KeaKAT Assignment

Figure 5. Selection of Keyphrases

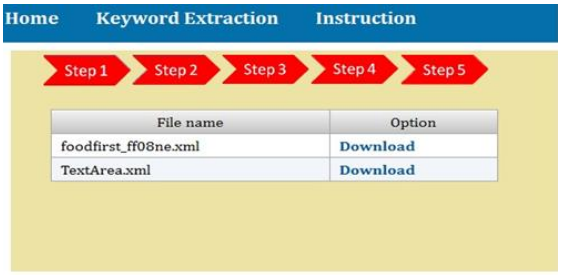

Figure 6. Download XML File

a) Step 4: In step 4, KeaKAT displays a result of keyphrase assignment. A user can select or reject the keyphrases assigned by the system. For example, if KeaKAT has assigned five keyphrases to a document and a user wants to select only four keyphrases from amongst them, then he can select keyphrases by selecting the check box and save them in an XML file as shown in Figure-5.

5) Generate XML file:

a) Step 5: All generated keyphrases are maintained in an XML file. The file is created for keyphrases of each document. A user can download the generated XML files as shown in Figure-6.

IV. Comparison with Existing Systems

A. Evaluation Metric

Quality is an essential characteristic of a software system, particularly web-based system [3], [2]. There is a wide range of metrics that has been proposed for quantifying the web quality attributes. However there is little consensus among them. These metrics focus on different aspects of web-based system or different quality characteristics. Usability is one of the important characteristics of quality [3], [2]. The focus of our research is on usability of online keyphrase assignment system.

International Standards Organization (ISO) defines usability as the degree to which specified users can achieve specified goals in a particular environment, with effectiveness, efficiency and satisfaction and in an acceptable way [12]. Moreover usability is a broad discipline based on the scientifically rigorous application of the observation, the measurement and principles of design useful for the creation of online systems

in order to bring to the final use of the system the ease of use, the speed of training, a high level of utility and a low level of discomfort. Usability problem causes a certain difficulty for an end user when interacting with a system. In[3], [2] the authors defines usability as a set of attributes that work on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users. The sub-characteristics of usability, defined there, are:

- understandability – attributes of software that work on the users' effort for recognizing the logical concept and its applicability;
- learnability – attributes of software that work on the users' effort for learning its application (for example, control, input, output);
- operability – attributes of software that work on the users' effort for operation and operation control;
- explicitness – attributes of software that work on the software product with regard to its status (progression bars, etc.);
- attractivity – attributes of software that work on the satisfaction of latent user desires and preferences, through services, behavior and presentation beyond actual demand;
- customisability – attributes of software that enable the software to be customized by the user to reduce the effort required for use and increase satisfaction with the software;
- clarity – attributes of software that work on the clarity of making the user aware of the functions it can perform;
- helpfulness – attributes of software that work on the availability of instructions for the user on how to interact with it; and
- user-friendliness – attributes of software that work on the users' satisfaction.

B. Usability Characteristics of Maui

In order to utilize the full features of Maui, one has to download, install and configure that, to be used as a desktop application. Moreover, the desktop application does not have any GUI features that can be helpful for a non-computing expert to assign keyphrases to documents. The GUI of the online demo version of Maui is simple and less interactive. The process of assigning keyphrases is supported by the option of text or PDF submission and vocabulary selection. But these details are far more technical for users belonging to non computer science domain.

The online version of Maui assigns keyphrases to documents from Agriculture and Physics domain. A user can upload vocabulary of its interest. It helps a user in assigning keyphrases to documents in the absence of any controlled vocabulary. In that case Maui utilizes Wikipedia to assign keyphrase. But Maui allows a user to assign keyphrases to a document based on the statistical model created by it and not based on the user provided training dataset in case a user wants to do that. However there is no progression bars to illustrate the status of keyphrase assignment. Poor help is available for document type and vocabulary. There is no facility to get a user feedback to improve upon the produced result.

C. Usability Characteristics of Agrotagger

The online automatic keyphrase assignment system: Agrotagger can only be used for classifying the agricultural domain documents. Like online demo version of Maui, Agrotagger adopts simple and less interactive GUI. Agrotagger offers a user to upload only PDF version of the document to assign keyphrase to them. Articles and text related to the working of Agrotagger are available online. However online version does not explain anything related to the usage of the system.

Agrotagger only offers keyphrase assignment related to the documents that belong to agriculture domain. As mentioned it assigns keyphrases from Agrotags which is formed from Agrovoc by selecting common and popular terms from it. It does not have the option of downloading the result set. A user cannot use his own training dataset to train the system. There is no option available to input user feedback in result set. No progress is shown through status bar.

D. Usability Characteristics of KeaKAT

The GUI of KeaKAT is enhanced and offers great usability support for users. The user can perform their tasks with an interactive system, that can take a user step by step to the system so that he can find it easy to use and know the status of his job. The help for using the system is separately provided to a user under *Instruction* tab. KeaKAT is available online so that user can access it from anywhere and at anytime.

The KeaKAT enables users to assign keyphrases to the document based on their own training dataset. A user can upload training documents one by one or in bulk. Agriculture and computing domain taxonomies are currently provided online by the system. However a user can upload a vocabulary of his own interest. Similarly test documents can be uploaded in a text and PDF file format one by one or in bulk of four documents. There is also *cut & past* facility for a test text/document. The result is presented to a user for its feedback before the final selection. The user can also download the result set in XML format so that it can be used by the user for other purposes.

E. Comparison

Maui and Agrotagger are designed for keyphrase assignment. They have good features, however they lack the usability characteristics. KeaKAT is particularly designed to guide and assist users of non-computing domain in keyphrase assignment; therefore its usability characteristics are better than the existing systems. We scale these systems on based of usability characteristics and assign the following categories: excellent, good, average, satisfactory, and poor as shown in table I. Moreover, the values stated in the table are based on the system study and data collected from users who participated voluntarily in the testing of documents using KeaKAT, Maui and Agrotagger.

V. Conclusion and Future Works

We have developed an online system on the basis of Kea++ and the extended refinement methodology. It can be used by

Table I
COMPARISON OF USABILITY CHARACTERISTICS

Characteristics	Maui	Agrotagger	KeaKAT
Understandability	Good	Good	Good
Learnability	Satisfactory	Satisfactory	Good
Operability	Average	Average	Good
Explicitness	Poor	Average	Good
Attractivity	Good	Average	Good
Customisability	Average	Poor	Good
Clarity	Poor	Poor	Good
Helpfulness	Satisfactory	Average	Good
User-friendliness	Poor	Average	Good

a user without the hassle of downloading and configuring. The system provides an attractive graphical user interface that guides a user in his operations and keeps him informed on the status of his jobs. KeaKAT facilitates users to easily assign keyphrases and generates an XML file of the result set which can be used for other purposes because XML is platform independent.

In future more taxonomies can be made online available in the system so that a user does not need to upload his own taxonomy. Currently the system only supports documents in English language. The system can be extended to assign refined and accurate keyphrases to documents in other languages as well. Comparison of KeaKAT with Agrotagger and Maui can be made on other different features .

REFERENCES

[1] A. T. Arampatzis, T. Tsoris, C. H. A. Koster, and T. P van der Weide. Phrase-based information retrieval. In *Proceedings of Information Processing and Management*, volume 34(6), pages 693–707, 1998.

[2] C. Calero, J. Ruiz, and M. Piattini. Classifying web metrics using the web quality model. *Online Information Review*, 29(3):227–248, 2005.

[3] Coral Calero, JuliÃ¡n Ruiz, and Mario Piattini. A web metrics survey using wqm. In *Web Engineering*, volume 3140 of *Lecture Notes in Computer Science*, pages 766–766. Springer Berlin / Heidelberg, 2004.

[4] I. Fatima, S. Khan, and K. Latif. Refinement methodology for automatic document alignment using taxonomy in digital libraries. In *IEEE International Conference on Semantic Computing*, pages 281–286, 2009.

[5] J. Feather and P. Sturges. International encyclopedia of information and library science, London & New York, 1996.

[6] G. Greg Oxton, J. Chmaj, and D. Kay. Perspectives on taxonomy, classification, structure and find-ability. Written for Consortium for Service Innovation.

[7] C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning, and E. Frank. Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems*, November 1999:81–104, 27(1-2).

[8] C. Gutwin, G. Paynter, I. H. Witten, C. Nevill-Manning, and E. Frank. Improving browsing in digital libraries with keyphrase indexes. Technical report, Department of Computer Science, University of Saskatchewan, Canada, 1998.

[9] Rabia Irfan. Extended refinement methodology for automatic keyphrase assignment. Master's thesis, School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan, 2012.

[10] S. Jones and M. Mahoui. Hierarchical document clustering using automatically extracted keyphrases. In *Proceedings of the 3rd International Asian Conference on Digital Libraries*, pages 113–120, 2000.

[11] Q. Li and Y.B. Wu. Identifying important concepts from medical documents. *Journal of biomedical informatics*, 36 (6):668–679, December 2006.

[12] C. Mariage, J. Vanderdonckt, and C. Pribeanu. State of the art of web usability guidelines. *The Handbook of Human Factors in Web Design*, pages 688–700, 1999.

[13] O. Medelyan. *Human-competitive automatic topic indexing*. PhD thesis, The University of Waikato, New Zealand, July, 2009.

[14] O. Medelyan and I. H. Witten. Thesaurus based automatic keyphrase indexing. In *Proceedings of the Joint Conference on Digital Libraries*, pages 296–297, 2006.

[15] NISO-ANSI. ANSI/NISO Z39.19 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies, 2005.

[16] P. Turney. Learning to extract keyphrases from text. Technical report, National Research Council Canada, 1999.

[17] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 254–255, 1999.