Issues and challenges of Research: Originality Based Document Classification and Information Extraction

Hussam Eddin Alfitouri Elgatait& Wan Mohd Fauzy
Centre for Instructional Technology and Multimedia
Universiti Sains Malaysia
Penang, Malaysia
hussamelgatait@hotmail.com, fauzy@usm.my

ABSTRACT— AS THE NUMBER OF PUBLICATIONS IN THE KNOWLEDGE DOMAIN IS STILL INCREASING, NATURAL LANGUAGE PROCESSING (NLP) AND TECHNIQUES FOR INFORMATION EXTRACTION FROM TEXTS BODY HAVE BEEN APPLIED WIDELY IN DIFFERENT FIELDS. IN ADDITION, THE SCIENTIFIC WRITING REQUIRES ARGUMENTS AND EVIDENCE, THE RESEARCHERS MUST BE CERTAIN THAT THE SOURCE OF INFORMATION IS RELIABLE, GENUINE AND VALID IN ORDER TO CONVINCE THE READER AND ENRICH THE KNOWLEDGE. THEREFORE, THIS PAPER HIGHLIGHTS THE CURRENT ISSUES AND CHALLENGES OF USING DIFFERENT INFORMATION EXTRACTION TECHNIQUES.

INDEX TERMS: INFORMATION EXTRACTION; SENTENCE CLASSIFICATION; COMPUTATIONAL LINGUISTIC; NATURAL LANGUAGE PROCESSING (NLP); KNOWLEDGE DOMAIN.

1 INTRODUCTION

Techniques for extract information's from texts body have become several and various. In addition to the above-mentioned, the targets of information extraction from texts have become urgent to decision-making; existing approaches focused on extract structured information from unstructured and identify some features from a collection of documents (i.e. location, relations between entities, author name, date...etc). In the same context, Information Extraction (IE) systems appear as tools to assist the information access, by extracting the parts that suitable to fill in a set of pre-defined output slots from the from a collection of documents [1]. Data extracted can be used to storage in database or used to presentation the data to user direct.

On the other hand, not all information available on the internet reliable (i.e. wikipedia, Yahoo, forums, etc) and these sites providing the information for learners and researcher in anytime and anywhere by provide them with full accessing and availability 24 hours a day, and seven days a week. And because the scientific writing require arguments, evidence and the evidence, so, the researchers must be certain that the source of this information it is judgements scientifically and Original sources, to convince the reader and enrichments the knowledge. in the domain of computer assisted language learning and natural language processing and information extraction technique is an important aspect of digital document, with information extraction the researchers/students can analyse and develop their trends of topics and relationship between the documents.

When the individual extract these metadata they can get high accuracy, but because we dealing with large number of these documents, progress of computer technology and information extraction technique demonstrated the efficiency and accuracy in this the domain[2], metadata extraction has become more interesting and researchers focus, we will focus on this study on the metadata extraction is refers here to extract metadata information from the references (i.e. Author, title, volume, etc) in the document and the relationship between these references in the candidate documents, (that address some problems such as errors in writing references and various styles of references) to proof the originality of the document and then prove the Information in Wikipedia it is reliability or not reliability, so, to is ensure that the information content matching with the content in the collection of documents, and then extract the metadata about this candidate document are to understanding the concept of free-text documents[3], text-to-text semantic[4], pattern-matching techniques [5], and latent semantic analysis approaches similarity existing approaches [6].

2 ISSUES AND CHALLENGES

Internet can be a great tool for learning and researching information, it's provides a useful information available to researchers and students. Ellsworth described the internet information as "a worldwide personal library," (Ellsworth 1995). Retrieval these information easy and safe time and effort by using the internet and available online any anytime anywhere. In addition, this information available around the world, students can learn, proves to be a good choice for supporting research and develop their search abilities. This kind of learning can be very useful for students as well as teachers. But some of these websites on the internet (i.e. wikipedia, E-Answers online, forums, etc) is cooperative effort of tens of thousands people, anyone can add and modify its content, these information may not constitute an acceptable source for a research paper.

Some of these websites does not publish original papers, Dictionary.com defined the Originality is "the quality or fact of being the product of individual creation that warrants copyright protec-

tion for a particular work regardless of novelty". The "original papers" refers to material (i.e. information, facts, ideas, etc) not all material published by reliable sources. This means not all material added to internet websites attributable to a reliable published source, so, must to find reliable sources about this topic, you should discover reliable source, and Because these websites is not the place to reliable information, and these materials that is challenged or likely to be challenged must be supported by a reliable source. Material for which no reliable source can be found is considered original research. Any cite you can show edit is not original research is to cite a reliable published source that contains the same material. Even with well-sourced material, if you use it out of context, or to advance a position not directly and explicitly supported by the source, you are engaging in original research.

Although the internet websites provided useful information for user in many fields (i.e. medicine, science, toxicology, cancer research and information drug, etc), and found that the depth these websites and the coverage was of a very high level, comparable in many cases, the coverage of databases, a doctor and is much better than Well-known reputation and national media. However, but still there many gaps when you use it, and some important issues about some information on products harmful and a source of great concern to areas such as medicine.

As in mentioned above, most of the people such as the researchers and students are looking through the Web for the information which are related to their research topics. The relative information of the research topics can be accessed from different resources (i.e. E-Journals; books published by university presses or published by respected publishing houses; magazines and mainstream newspapers). The researchers are focusing on the rightness of the gained information in order to consider the source as a reference. So many researchers are referring to the articles that can be accessed from the e-Journals (i.e. IEEE, ACM …) because these e-Journals have restrictions on the published articles. On the other hand, most of information resources do not have any restriction on their published articles such as; wikipedia, forums, etc. Consequently, invalid information may be published and then these recourses can't be considered as references. Based on these issues there is need to a method that can measure the information validity of these resources is in demand in order to be considered as a reference, so, we need further investigate for prove the information in the internet if reliable, genuine and valid or not in order to convince the reader and enrich the knowledge.

A number of researchers [7-11] addressed the drawbacks of the current detection tools that are:

- Lacking to distinguish correctly cited text from plagiarized text.
- lacking to involve the books category.
- lacking to detect plagiarized words based ideas.
- Lacking to process textual images for similarity checks.

However, indicated the way that most of detection methods follow for analysing analysis the document originality, which shows that although these tools provide excellent service in detecting matching text between documents, but it's still lacking to distinguish correctly cited text from plagiarized text is one of the serious drawbacks of these tools. That is why utilize new technique is still needed.

3 Motivation

As demonstrated in this document, the numbering for sections upper case Arabic numerals, then upper case Arabic numerals, separated by periods. Initial paragraphs after the section title are not indented. Only the initial, introductory paragraph has a drop cap.

4 Literature Review

Current trend detection methods fall into two categories: fully-automatic and semi-automatic. In fully-automatic approach, a list of emerging topics is developed afterwards researcher pursues these topics and the evidence to determine which truly emerging trends are. Semi-automatic approach on the other hand, requires user to input a topic then it provides the user with the evidence whether the input topic is truly emerging [12]. Most of the existing trend detection systems focus on keyword matching, statistical techniques and link analysis. Currently there are emerging efforts to employ Semantic Web technologies to provide enhance information search and retrieval mechanisms. The Semantic Web has been desired as an extension of the current Web, which makes a well-defined meaning of information to enable a better connection between computers and people to work together [13]. Ontology is used to visualize knowledge on the Semantic Web. Recently, ontology is a formal representation of a set of concepts within a domain into a machine-readable format that is also understandable by humans, consisting of entities, attributes and relationships [14].

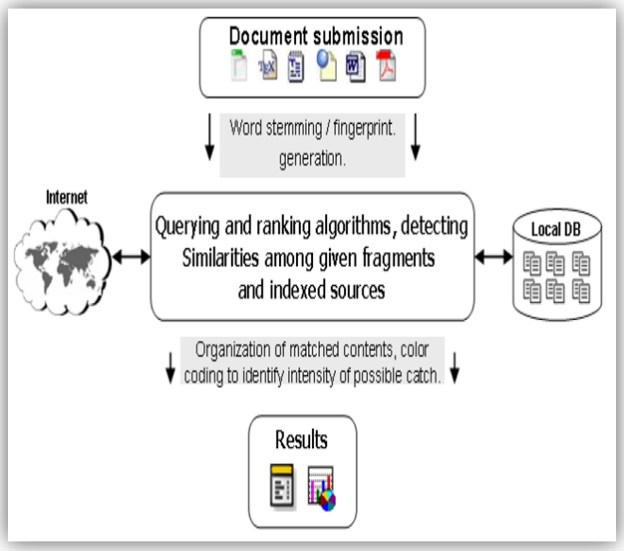

Figure 1: A generic structure of document source comparison based on plagiarism detection system [11]

Figure 1 presents the traditional method for comparing two or more documents and to reason about degree of similarity between them, it is needed to assign numeric value, so called, similarity score to each document. This score can be based on different metrics. There are many parameters and aspects in the document which can be used as metrics.

According to Bahrami in [7] highlights the importance of the physical components and sub-components of the e-document to be converted to an XML e-document that can be used to match similarity or document originality. Bahrami indicated the main issues in linking the conversion process that justify the issue in identifying the components and sub-components of the e-document. Bahrami adopted a methodology, blocking scheme to identify semi-structured e-documents' components and sub-components. The methodology was applied on 50 semi-structured e-documents of a specific type and the identification of their components and sub-components were achieved with 100% accuracy.

Deeper, Mahdavi et al. (2009) reports the prime issues in analyzing and interpreting document contents. Trend detection in scientific publication retrieval systems helps scholars to find relevant, new and popular special areas by visualizing the trend of input topic. Authors aimed to describe the difficulties of previous researches to obtain a suitable classification model for research topics that it's cited by other researchers. Mahdavi et al. developed a technique that combines both of semantic components and ontology classification to provide an advance functions for trend detection in the context of scholarly Semantic Web system (SSWeb) [8].

Shaparenko and Joachims (2009) address the main issues in specifying the cited articles that most of research papers referred to during their research and how text mining provide readers with automatic methods for quickly finding the key ideas in individual documents and whole corpora. Shaparenko and Joachims proposed a statistically well-founded method that aimed to determine and analyze the original ideas that a document contributes to a corpus, focusing on self-referential diachronic corpora such as research publications, blogs, email, and news articles. This model works on indicating the original content through a combination of impact and novelty, and it can be used to identify the most original passages in a document. Shaparenko and Joachims evaluated the developed model among synthetic and real data. The obtained result showed that the model outperforms a heuristic baseline method [9].

Finally, Van et al. (2008) reports the importance of providing a way for detecting that original author for a certain document, which consider to be an important role in specific domains. It's also showing how the current methods in identifying and detecting authors work are lacking to describe and generalize the full features for this class of documents. Van et al. proposed a new approach for detecting false documents using a document signature obtained from its intrinsic features: bounding boxes of connected components are used as a signature. The work process for this approach lied on utilizing the optimal document alignment to build a model signature that can be used to compute the probability of a new document being an original one. Preliminary evaluation shows that the method is able to reliably detect faked documents.

The process starts with painting the original black pixels from the document in blue; and then, the black pixels from the original document will be painted in red. The procedure of this approach will indicate the differences in the red and blue pixels. It can be seen that the copying process seems to move blocks up and down: left part the copy is too far down, the middle parts quite well and the right part is again to far down [10].

5 Conclusion

This paper aimed to put the sight on the current challenges and issues faced by different academicals institutions towards using natural language processing for checking research originality by classifying and extracting information inside the research papers. Prior researches on the current techniques were addressed to verify these issues.

6 References

[1] [1] J. Turmo, et al., "Adaptive information extraction," ACM Comput. Surv., vol. 38, p. 4, 2006.

[2] [2] Z. Ni and H. Xu, "Automatic Citation Metadata Extraction Using Hidden Markov Models," presented at the Proceedings of the 2009 First IEEE International Conference on Information Science and Engineering, 2009.

[3] [3] R. Williams and J. Nash, "Computer-Based Assessment: From Objective Tests to Automated Essay Grading. Now for Automated Essay Writing?," in Information Systems: Modeling, Development, and Integration. vol. 20, J. Yang, et al., Eds., ed: Springer Berlin Heidelberg, 2009, pp. 214-221.

[4] [4] R. Mihalcea, et al., "Corpus-based and knowledge-based measures of text semantic similarity," presented at the Proceedings of the 21st national conference on Artificial intelligence - Volume 1, Boston, Massachusetts, 2006.

[5] [5] N. K. Nikitas, "Computer Assisted Assessment (CAA) of Free-Text: Literature Review and the Specification of an Alternative CAA System," 2010, pp. 116-118.

[6] [6] P. Dessus, "An Overview of LSA-Based Systems for Supporting Learning and Teaching," presented at the Proceeding of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, 2009.

[7] [7] A. A. Bahrami, "A Blocking Scheme for Identification of Components and Sub-Components of Semi-Structured E-Documents," in Fifth International Conference on Information Technology: New Generations, NW-US, 2008, pp. 943-948.

[8] [8] F. Mahdavi, et al., "Semi-Automatic Trend Detection in Scholarly Repository Using Semantic Approach," World Academy of Science, Engineering and Technology, vol. 5, pp. 224-226, 2009.

[9] [9] B. Shaparenko and T. Joachims, "Identifying the original contribution of a document via language modeling," Machine Learning and Knowledge Discovery in Databases, pp. 350-365, 2009.

[10] [10] J. Van, et al., "Document signature using intrinsic features for counterfeit detection," Computational Forensics, pp. 47-57, 2008.

[11] [11] J. Jaya, "Plagiarism Detection Techniques," Cochin University Of Science And Technology, 2007.

[12] [12] B. Aleman-Meza, et al., "Template based semantic similarity for security applications," Intelligence and Security Informatics, pp. 621-622, 2005.

[13] [13] K. Verma, et al., "Meteor-s wsdi: A scalable p2p infrastructure of registries for semantic publication and discovery of web services," Information Technology and Management, vol. 6, pp. 17-39, 2005.

[14] [14] M. A. Ismail, et al., "Semantic support environment for research activity," Journal of US-CHINA Education Review, vol. 5, pp. 36-51, 2008.