Measuring the Impact of Arabic Scientific Publication: Challenges and Proposed Solution

Raad Alturki [1]

[1]*ralturki@ccis.imamu.edu.sa*
Department of Computer Science, Al-Imam Mohammad Ibn Saud Islamic University,
P.O.Box: 5701, Riyadh:11432 (Saudi Arabia)

Introduction

Citation Indices are very useful tools that were firstly used to help finding articles easily and then, used to provide information about research output. They can be used as indicator to measure research performance, provide information about trends in research and compare and rank the research output of countries, institutes and authors. It is well known that English is the universal language for science and technology and that have resulted in having many citation indices like Web of Science (Formerly ISI) and SCOPUS. It has been reported in the literature that such Indices overlook and hide publications in other languages (van Leeuwen et al., 2001) and that -with other reasons- have resulted in having indices for other languages like Chinese, Portuguese and Korean. Arabic publications is one of the least represented in the scientific community despite its been spoken by more than 200 million which makes it the fifth spoken language in the world (Gordon Jr., 2005). This work investigates the possibility of making a Citation Index for Arabic literature and addresses the challenges associated with that. This is supported by initial implementation of web based Arabic Citation Index (ACI).

Challenges

This section discusses challenges associated with non-English citation indices with special focus on the one dealing with Arabic literature. In order to have citation index for any language, it is very important to make it integrate with other English-based indices. Non-English citation indices should be able to read citations from other indices in order to see how any article or language is impacting the scientific community. This raises some issues of how to make cross languages referencing; if an article written in Chinese has cited other article in Korean, how the Chinese/Korean indices will identify this citation. This problem is not easy to be solved unless if there is a well established standardization for citations which allows identifying any article in any language. Such identifier should be unique across the globe and can be used in every citation. Luckily, Digital Object Identifier (DOI) can be used to serve this purpose while the adoption of using DOI in referencing is not yet being very popular as citation styles are still not considering that as part of the cited article. Having DOI as a compulsory in each citation style makes it easier for articles to be identified, then cited and discovered in citation indices across languages.

Unfortunately, there is no enough information about the scientific contribution written in Arabic. One of the most accurate information we found is the number of periodicals that have ISSN. According to a report by ISSN foundation, in 2012 there were 4489 new periodical record in Arabic which makes it the 26th most registered language in the world. The ISSN records do not represent only scientific journals but it registers any types of periodical. Also, there is a report by Thomson Reuters about the contribution of Arab countries recorded in their databases. The report shows that the number of scientific documents produced in those countries is around 13,574 in 2008 (Adams et al., 2011) where most of the written articles are in English. In fact, there are many journals written in Arabic that are not well recognized in the internet and digital libraries. We have noticed that Arabic scientific journals are still focusing on publishing printed format with no much focus on the electronic version.

In reality, there are some digital libraries that aggregate articles of major Arabic journals and provide electronic versions of such articles. However, having seen some of the main digital libraries and aggregators in Arabic, we still believe such aggregators have some issues as they provide the articles as scanned documents that cannot be indexed automatically. Also, such digital libraries do not have the full bibliographic information like title, abstract, authors, year of publishing, publisher name, volume, ISSN and list of references. Having bibliographic information is vital for building any citation index as they are the raw data to draw the relationship between article and scientific work in term of citations. If bibliographic information is not available for any reason, the PDF electronic version of the article could be used to extract the bibliographic information. Extracting such information from any electronic file can be done with some challenges if the article is saved as text rather than picture. The process becomes very

sophisticated if article is saved as picture where scanning should be done properly. Then Arabic text recognition algorithm should be used to recognize text used when current algorithms in Arabic are not reliable and accuracy rate is low.

Additional challenge in working with Arabic literature is the lack of standardization of the structure and the location of different section in articles. Any software that scan or parse the paper will make some assumptions of the location of the title, authors and abstract. Google scholar software that extract bibliographic information from files directly without having bibliographic information assumes that first line is the title which is written in large font. It has been stated in a study of Arabic journals that "instructions to authors" are generative and are not precise enough (Alkholaifi, 2001). That results in having different interpretations of instructions specially in using referencing style. Variations in formatting could happen at different places of the article, including authors' names, authors' salutation (Dr, professor), availability of abstract and list of references. List of references can be written in mixture of two languages at the same time (Arabic and English) which makes extraction harder. The extraction program should be able to work with different languages at the same time and be able to differentiate between different citing styles.

Extracted Information from article could include errors that can be stored in the index. The program should be aware of such errors and correct them before storing. Detecting errors is not an easy task as it should understand the context of the information. Names sometimes could be recognized as error or misspelled words as some names could have different variations or do not have a direct meaning especially if the name is not Arabic. After the information about any specific word is stored in the index, a query can be done to find a specific article or articles in certain subject. For this reason, search query should be able to consider all possible errors that user might have done when entering the keywords beside the stemming and lemmatization process that happens at indexing phase. In fact, there are several Arabic spelling correction techniques (Manning et al., 2006; Attia et al., 2012; Larkey et al., 2002; Rytting et al., 2011; Shaalan et al., 2012). Using such techniques will be of great important in implementing any Arabic based citation index. These techniques in Arabic are similar to other languages with few differences include the morphological analysis and context understanding of the language where Arabic language is complex in comparison to English.

The proposed system

The overall architecture of the system is shown in Figure 1 where it shows the five main components: Crawler, Parser, Matcher, Database and User Interface. This architecture is inspired by the typical design of search engines as they share similar concepts. One major difference between the two systems is that citation indices use citations as way to rank and measure the impact of an article whereas search engines normally uses the links and other metrics as a way to rank sites and documents.

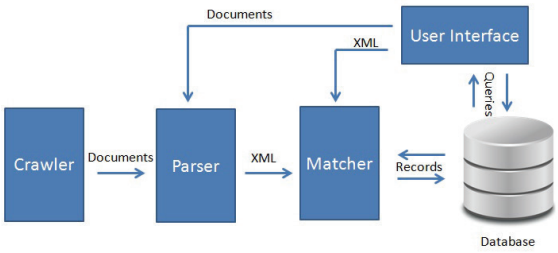

Figure 1. The proposed Architecture of ACI.

References:

Adams, J., King, C., Pendlebury, D., Hook, D. & Wilsdon, J. (2011). Global research report. Middle East, *Evidence*, Thompson Reuters.

Alkholaifi, M. (2001). Documenting citations: an analytical study of publishing policy in some journals. *Journal of King Fahd National Library* vol. 6.

Attia, M., Pecina, P., Samih, Y., Shaalan, K. F., & van Genabith, J. (2012). Improved Spelling Error Detection and Correction for Arabic. *Proc. COLING,* 103-112.

Gordon Jr, R. G. (2005). Ethnologue: Languages of the World, Dallas, Tex.: SIL International. *Online version*: http://www. ethnologue. com.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. *Proc. 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 275-282.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge: Cambridge University Press.

Rytting, C. A., Zajic, D. M., Rodrigues, P., Wayland, S. C., Hettick, C., Buckwalter, T., & Blake, C. C. (2011). Spelling correction for dialectal Arabic dictionary lookup. *ACM Transactions on Asian Language Information Processing* (TALIP), *10*(1), 3.

Shaalan, K. F., Attia, M., Pecina, P., Samih, Y., & van Genabith, J. (2012). Arabic Word Generation and Modelling for Spell Checking. *Proc. LREC*, 719-725.

Van Leeuwen, T., Moed, H., Tijssen, R., Visser, M., & Van Raan, A. (2001). Language biases in the coverage of the Science Citation Index and its consequencesfor international comparisons of national research performance. *Scientometrics*, *51*(1), 335-346.