Mapping the Question Answering Domain

Mohan John Blooma, Alton Yeow Kuan Chua, Dion Hoe-Lian Goh

Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore 637718
{bl0002hn, AltonChua, ASHLGoh}@ntu.edu.sg

Abstract. We present a trend analysis of the question answering (QA) domain. Bibliometric mapping was used to sketch the boundary of the domain by uncovering the topics central to and peripheral to QA research in the new millennium. This paper visualizes the evolution of concepts in the QA domain by studying the dynamics of the QA research during the periods 2000 – 2003 and 2004 – 2007. It was found that question classification, answer extraction, information retrieval, user interface, performance evaluation, web, & natural language were the main topics in current QA research.

Keywords: bibliometric mapping, co-word analysis, question answering

Question Answering (QA), a pertinent branch of information retrieval is an emerging research field with roots traced back to 1960s. Studies in QA research aim in building intelligent systems that can provide succinct answers to questions constructed in natural language. As the literature in QA research evolves dynamically and proliferates in diverging research directions, the task of charting the intellectual structure of the domain becomes increasingly challenging. This study aims to mark the perimeters of the research trends in the QA domain using widely accepted bibliometric mapping tools [3, 4].

Papers containing the term "question answering" in the title were used as the selection criteria to download from ACM digital library. The documents were restricted to only journals and conferences between 2000 and 2007. The title, abstracts and keywords were extracted. Connexor was used to extract noun phrases and TEXTSTAT2 was used for identifying most frequent phrases. Co-word analysis was performed on phrases in this study. Co-word analysis is built on the assumption that a paper's keywords constitute an adequate description of its content. It is based on the nature of words, which are the important carrier of scientific concepts, idea and knowledge. Two keywords co-occurring within the same paper indicates a possible link between the topics to which they refer. The presence of many co-occurrences around the same word or pair of words highlights a locus of considerable association within papers that may correspond to a research theme. Pajek, a freeware program for visualization developed by the University of Ljubljana, is used in this study for mapping. The QA domain in this study is visualized using the algorithm of Kamada & Kawai as it is available in Pajek. 30 most frequent words to map the domain.

Sketching the boundary of QA domain: 2000 – 2007: The network map illustrated that the core concepts related to QA are "information retrieval", "question type", "answer", and "performance" forming the inner circle. "Information search",

"experimentation", "information storage", "algorithm", "design", "natural language" and "web" are the phrases that form the outer circle. Hence it could be concluded that the phrases in the inner circle are highly correlated terms in the QA domain. Analysis of the period 2000-2003: It was evident that during the period 2000 – 2003, QA research were centred on "question classification", "web", "performance" studies, "information retrieval", "answer" extraction techniques, and "search engines". However, the concepts like "natural language", "information systems", "knowledge annotation", "user", & "interface" were plotted as distant disciplines to "question answering". "Ontology", "predictive annotation" and "machine learning" are other new areas that were mapped in this network analysis. Analysis of the period 2004-2007: It was found that phrases "question type", "answer", "information search" and "information retrieval" were highly co-related terms to QA. "Automatic", "syntactic" "pattern" were terms that were unique to this period indicating that they are the emerging areas in the QA domain.

From the above results, it could be inferred that information retrieval, information storage and search are the subject areas that are highly co-occurring with the phrase question answering. This finding is in evidence to the universal definition of QA that it is a type of information retrieval. "Question type" is a very closely related concept in QA domain and hence Question classification studies are a major research trend in QA domain [1]. The third concept that has been focused in the QA research since 2000 is "performance". The fourth concept is inevitably the answer extraction [2]. From a bird's eye view of the QA domain obtained from this study it could be inferred that QA domain is a multi disciplinary domain.

By conducting a co–word analysis for two time periods 2000 – 2003 and 2004 – 2007, it could be concluded that computational linguistics and artificial intelligence are the emerging trends during the last five years. Enhancing the performance, user studies, and enriching the algorithms have gained higher similarity rather than studies on question classification and answer extraction during the recent years. Results of this study educe the multidisciplinary nature of QA. It also paves way to the future of QA and predicts that there will be a tremendous growth in research related to the user interaction and computer linguistics areas other than its paternal domain of information retrieval.

References

1. Blooma, M.J., Chua, A., Goh, D.: A predictive framework for retrieving the best answer. In the proceedings of the ACM Symposium on applied computing, 1107-1111.
2. Blooma, M.J., Chua, A., Goh, D.: Applying question classification to yahoo answer. In the proceedings of the first IEEE International Conference on the Applications of Digital Information and Web Technologies (2008).
3. Callon, M., Courtial, J. P., Turner, W., & Brain, S.: From translations to problematic networks. An introduction to co-word analysis. Social Science Information, 22 (2), (1983) 191 – 235.
4. Janssens, F., Leta, J., Glanzel, W., & Moor, B. D.: Towards mapping library and information science. Information Processing and Management 42, (2006) 1614–1642.