Review of Knowledge Extraction from Email: The Boundaries of Natural Language Processing Techniques

T.W.Jackson
Information Science
Research School of Informatics
Loughborough University
Loughborough, LE11 3TU, UK

T.W.Jackson@lboro.ac.uk

S.Tedmori
Computer Science Department
Princess Sumaya University for Technology, Al-Jubaiha 11941
Jordan

S.Tedmori@psut.edu

C.J.Hinde and A.I.Bani-Hani
Computer Science
Research School of Informatics
Loughborough University
Loughborough, LE11 3TU, UK

{C.J.Hinde, A.I.Bani-Hani}@lboro.ac.uk

***Abstract*—The aim of this research is to determine if natural language processing techniques can be used to fully automate the extraction of knowledge from emails. The paper reviews the four generations of building systems to share knowledge and highlights the challenges faced by all. The paper shows that although the f-measure results are world leading for this study, there is still a requirement for user intervention to enable the system to be accurate enough to be of use to an organisation.**

I. INTRODUCTION

Over the last several decades, many reports [1], [2], [3], [4], [5] have indicated that people searching for information prefer to consult other people, rather than to use on-line or off-line manuals. Allen [5] found that engineers and scientists were roughly five times more likely to consult individuals rather than impersonal sources such as a database or file cabinet for information. In spite of the advancements in computing and communications technology, this tendency still holds; people remain the most valued and used source for knowledge[6], [7].

Unfortunately, finding individuals with the required expertise can be extremely expensive [8], [9], as it is time consuming and can interrupt the work of multiple persons. A common problem with many businesses today, large and small, is the difficulty associated with identifying where the knowledge lies. A lot of data and information generated and knowledge gained from projects reside in the minds of employees. Therefore the key problem is, how do you discover who possesses the knowledge sought?

In the search for the solution, information systems have been identified as key players with regards to their ability to connect people to people to enable them to share their expertise and collaborate with each other [10], [11], [12], [13]. Thus, the solution is not to attempt to archive all employees' knowledge, but to link questions to answers or to knowledgeable people, who can help find the answers sought [14]. This has led to the interest in systems, which help connect people to others that can help them solve their problems, answer their questions, and work collaboratively.

Cross et al. [15] reviewed [16], [17], [18], [19], [20], [21], [22], [23], [24], and summarises the benefits of seeking information from other people. These benefits include:

- provision of solutions to problems;
- provision of answers to questions;
- provision of pointers to others that might know the answer
- provision of pointers to other useful sources;
- engagement in interaction that helps shape the dimension of the problem space;
- psychological benefits (e.g. confidence, assurance);
- social benefits (e.g. social approval for decisions, actions);
- improvement in the effectiveness with which a person advances their knowledge in new and often diverse social contexts;
- improvement in efficiency (e.g. reduction in time wasted pursuing other avenues); and
- legitimation of decisions.

Cross [15] identifies five categories that these benefits fall under: (1) solutions (know what and know how); (2) meta-knowledge (pointers to databases or other people); (3) problem reformulation; (4) validation of plans or solutions; and (5) legitimation from contact with a respected person. It has been recognised that the idea of connecting people to people is a way forward, yet from a natural language processing viewpoint what has been attempted before and what are the limitations of the current systems.

This paper reviews the expert finding approaches and discusses the natural language processing (NPL) techniques used to extract knowledge from email, including the one developed by the

authors. It concludes by reflecting on the current f-measure scores for knowledge extraction and the role of the user in any knowledge location system.

II. EXPERT FINDING APPROACHES

Various approaches to expertise location have been developed and implemented to link expertise seekers with internal experts. The first generation of such systems sprung out of the use of helpdesks as formal sources of knowledge, and comprised knowledge directories and expert databases. Microsoft's SPUD project, Hewlett-Packard's CONNEX KM system, and the SAGE expert finder are key examples of this genre. Generally expert databases have 'Yellow Pages' interfaces representing electronic directories of experts linked to their main areas of expertise. Such directories are based on expert profiles which must be maintained by experts on a voluntary basis. The key advantages of such directories include conveniently connecting those employees inadequately tapped into social and knowledge networks with relevant experts. However such approaches also suffer from significant shortcomings. Unless employees regularly update their profiles, the profiles lose accuracy and no longer reflect reality. Yet employees are notorious for neglecting to update such profiles as such duties are often considered onerous and low priority [25]. Employees may not wish to provide expertise. Overall, when large numbers of employees are registered and profiles are inaccurate, credibility is rapidly lost in such systems which are increasingly ignored by knowledge seekers, who instead rely on social networks or other methods [9]. In addition, expertise descriptions are usually incomplete and general, in contrast with the expert-related queries that are usually fine-grained and specific, and replete with various qualitative requirements [25].

In the second generation of expertise locators, companies took advantage of personal web pages where employees could advertise expertise internally or externally. Such pages are designed according to corporate templates or personal design, and are usually made accessible via the World Wide Web or corporate intranets. The convenience of web site creation and update, web site retrieval and access, and sophisticated search engines, are key advantages of this approach. However, employees may lack the motivation, time or technical expertise to develop or update their profiles, which rapidly lose accuracy and credibility and the capacity to meet expert location needs [25]. In addition, as noted by Yimam-Seid and Kobsa [25], employee use of search engines for locating an expert's web page may be ineffective since such a process is based on a simple keyword matching task which does not always yield the most relevant experts' web pages. The search activity can also be very time consuming when a high number of hits is returned and an employee must then systematically or randomly attempt to choose and explore the listed link(s). As Yimam-Seid and Kobsa have observed for this approach, knowledge seekers are allocated significant and often onerous responsibility for finding relevant experts ([25]. The second generation of approaches also included the development of more dynamic expert databases. *Answer Garden* [26], [27], which is a question-answering system, maintains a database of frequently asked questions and answers. When the system does not find required information in the database, an end-user may ask the question of the system. *Answer Garden* then routes the question to the corresponding experts. However, it is not clear with this approach how the system identifies experts and, in particular, whether experts have nominated their own areas and levels of expertise.

The third generation of approaches relies primarily on secondary sources for expert identification. For example, the web application *Expertise Browser* [28], studies browsing patterns/activities in order to identify experts. With this application, if the user knows a particular expert, the user can ask the system to reveal the browsing path of that expert, relevant to the user's query. Among other disadvantages, if an employee does not know an expert, the user must ask the system to identify one or more experts. The employee must then scan the browsing paths of the identified experts for possibly useful links, which can be a very time consuming process. Furthermore, it is likely that browsing reveals interests rather than expertise [25]. The monitoring of browsing patterns clearly involves privacy issues that such systems fail to address. Other secondary-source approaches utilise message board discussions as indicators of expertise. For example, *ContactFinder* [29] is a research prototype that reviews messages posted on message boards. *ContactFinder* analyses subject areas from messages and links them to the names of experts who wrote the messages. It provides users seeking experts with expert referrals when user questions match expert's earlier postings. All such approaches infer experts from secondary sources but do not allow experts to confirm such inferences.

A recently recognised socially based approach is the use of social networks which provide a complex social structure for the development of social capital and the connection of novices and experts [11]. In a study conducted by [7], while some people were difficult to access, they were still judged to be valuable sources of help. The use of a social network to locate expertise has become popular because colleagues are often physically available, are personal friends, or are known to be experts on the topic. However, there is no guarantee that a genuine expert will be consulted, as users may choose to consult a moderately knowledgeable person, a person with whom a good relationship exists, a proximate employee, or a quickly located employee, simply as that person is within the expertise seeker's social network. With this approach, low quality expertise may be systematically input into an organisation where it is quickly applied. Automated social network approaches such as *Referral Web* suffer from similar concerns.

The fourth generation may include one or more the above approaches together with natural language processing and artificial intelligence techniques in order to analyse stored knowledge, seeking to identify expertise and experts [25], [30], [31]. A forerunner of such systems was *Expert Locator* which returns pointers to research groups in response to natural language queries on reports and web pages [32]. A second example is *Expert Finder* [33] which considers self-published documents containing the topic keyword, and the frequency of the person named near the same topic keyword in non-self-published documents, in order to produce expertise scores and ranks. In 1993 Schwartz and Wood first attempted to utilise e-mail messages, known to be heavily knowledge-based, to deduce shared-interest relationships between employees. In 2001, following other experts' promising attempts, Sihn & Heeren implemented *XpertFinder,* the first substantial attempt to exploit the knowledge-based content of e-mail messages by employing

technology to analyse message content. More recently Google Mail have use similar techniques to scan email content whilst reading messages on-line, to extract key phrases that can then be matched with specific marketing adverts that appear to the right hand side of the browser. This is more a case of just-in-time knowledge that could be extremely useful to employees if, for example, they were writing reports and the application would mine for keywords and link the user to existing written material or experts to aid in the report writing task.

The major drawback of many of the fourth generation approaches is that output such as potential expert listings is unordered when presented to a user seeking experts, requiring significant user effort to identify the best expert. Such systems identify experts by textual analysis but rarely support expert selection by users. In addition, such systems fail to present the varying degrees (or levels) of expertise that people possess and tend to assume a single level of expertise. It is thus entirely the user's responsibility to systematically process the returned results in order to identify the most suitable experts for answering specific queries. Techniques employed to build the fourth generation expertise profiles should be advanced to ensure that the textual fragments analysed accurately convey employees' expertise. To date, the automated techniques have been inadequate because they cannot distinguish between what is important and what is not important in identifying an expert. In addition, the system should be able to match user needs with expertise profiles by using appropriate retrieval techniques, ensuring that relevant experts are not overlooked and that less relevant experts are not overburdened with inappropriate queries.

This abbreviated evolutionary review of expertise locator systems has highlighted the need for new expert locator systems with enhanced information retrieval techniques that provide user friendly expertise seeking techniques and high levels of accuracy in identifying relevant experts. In the next section, we summarise the techniques that have been used to extract key phrases and then discuss the latest attempts by the authors to improve upon the techniques to enhance the accuracy of the key phrases extracted and the ranking of their importance according to a user's expertise.

III. KEY PHRASE EXTRACTION

Numerous papers explore the task of producing a document summary by extracting key sentences from the document [34], [35], [36], [37], [38], [39], [40], [41], [42], [43].

The two main techniques are domain dependent and domain independent. Domain dependent techniques employ machine learning and require a collection of documents with key phrases already attached, for training purposes. Furthermore, the techniques (both domain dependent and domain independent) are related to linguistics and/or use pure statistical methods. A number of applications have been developed using such techniques. A full discussion of existing approaches, together with their merits and pitfalls, is provided in [44].

There are many weaknesses with current approaches to automatic key phrase identification, several of which are discussed here to illustrate the issues. First, the extraction of noun phrases from a passage of text is common to all such approaches [43], [45]. However, a disadvantage of the noun extraction approach is that, despite the application of filters, many extracted key phrases are common words likely to occur in numerous e-mails in many contexts. Therefore, it is important to distinguish between more general nouns and nouns more likely to comprise of key phrases. Second, Hulth [45] pinpoints two common drawbacks with existing algorithms, such as KEA. The first drawback is that the number of words in a key phrase is limited to three. The second drawback is that the user must state the number of keywords to extract from each document [45].

In the attempt to push the boundaries of key phrase extraction, work undertaken by the authors aimed to enable end-users to locate employees who may possess specialised knowledge that users seek. The underlying technical challenge in utilising e-mail message content for expert identification is the extraction of key phrases indicative of sender skills and experience.

In developing the new system the Natural Language ToolKit (NLTK) was employed to build a key phrase extraction "engine". NLTK comprises a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language. The completed key phrase extractor was then embedded within EKE - an Email Knowledge Extraction process based on two stages.

The first stage involves a training process which enables the creation of a speech-tagging model for tagging parts-of-speech (POS) within an e-mail message. The second stage involves the extraction of key phrases from e-mail messages with the help of the speech-tagging model.

Figure 1 depicts how the EKE system analyses e-mail messages to identify experts. Once a message is sent by a user (step 1), the body of the message is captured by EKE. EKE's key phrase extraction engine will parse the body of the email seeking appropriate key phrases that might represent the user's expertise (step 2). This process is fully automated and takes only milliseconds to complete, and is so far transparent to both sender and receiver. It is possible that key phrases will not be identified by the key phrase extraction engine as the message may not contain any text suggesting key phrases, or the message contains key phrases that were not detected. In such cases, EKE will not require any action from the user whose work activities will therefore remain uninterrupted.

In step 3, if the engine identifies key phrases the user is requested to rank the extracted key phrase using a scale of 1 - 4, to denote level of user expertise in the corresponding field. The rankings 1 – 4 represent basic knowledge, working knowledge, expert knowledge, or not applicable. The four point categorisation scale was devised because a seeker of knowledge should be forewarned that a self-nominated expert may lack an expected capability. The knowledge seeker can then decide whether to proceed to contact such an expert for help. In Figure 1, "Questionnaire", "Semantics", "Casino" and "Online database" are examples of the key phases that have been extracted from the body of a message. On average very few key phrases are extracted from a message because generally, according to our development tests and pilot studies, there are few key-phrases contained within any one e-mail message. Therefore typically a user is not unduly delayed by the

key phrase expertise categorisation process. Once categorised (for example in Figure 1, "Questionnaire" may be categorised as basic knowledge, "Semantics" as expert knowledge, and so on), key phrases are stored in an expertise profile database (excluding key phrases categorised as "not applicable"). The user can access and edit his/her expert profile at anytime (step 4). The key phrases that are stored in the expertise profile database are also made available to other employees within the organisation, to enable them to locate relevant experts by querying the database (step 5).

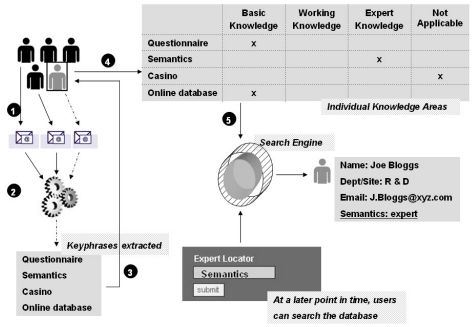

Figure 1 – Overview of the E-mail Knowledge Extraction System [44]

The EKE system has significant advantages compared with other e-mail key phrase extraction systems, not all of which perform steps 3 and 4. The present system gains accuracy by requiring a user in steps 3 and 4 to rank his or her level of expertise for a particular key phrase. Most existing systems attempt to rank experts automatically rather than consulting users for their perceptions of their level of expertise. Such systems are likely to be less successful at accurately identifying expertise levels as they do not capture employee knowledge of their own expertise. The above approach has been trialled at Loughborough University as mentioned in the Research Design section, and shown to be effective in correctly identifying experts [44]. However, it is important to note that this system uses a hybrid approach of NLP and user intervention to determine the usefulness of the key phrases extracted. User intervention was introduced after the results of the NLP system were not accurate enough to fully automate the process. The next section reviews the results of the NLP system without user intervention, which leads to a discussion about the boundaries of NLP in key phrase extraction.

IV. RESULTS AND BOUNDARIES OF NLP

The Natural Language ToolKit system developed by the authors was tested on a number of corpuses (not the full EKE system which includes user intervention).

- Corpus 1 - Emails from various academic domains; Size 45
- Corpus 2 - Total office solutions organisation; Size 19
- Corpus 3 – Enron; Size 50

The *sampling units* were collected from subjects from different backgrounds (people with English as their first language and people who can communicate in English, but is not their first language). All subjects belong to the age group 24-60. All the *sampling units* were outgoing mail. The authors believe that *sampling units* are representative of typical messages that are sent out in institutional and corporate environments. The *sampling units* of the *sample*, Corpus 1, were collated from various academic disciplines (computer science, information science, building and construction engineering). The *sampling units* of the second *sample,* Corpus 2, are specific to one employee from a large supplier of total office solutions in the UK & Ireland, which for confidentiality reasons in is referred to as Employee E from Company XYZ. The *sampling units* of the final sample, Corpus 3, are collated from the Enron email dataset, which is freely available on the net.

The f-measure, a widely used performance measure in information retrieval, was used to measure the system and is defined as:

$$f-measure = \frac{2\times precision \times recall}{precision + recall}$$

where *precision* is the estimate of the probability that if a given system outputs a phrase as a key phrase, then it is truly a key phrase and *recall* is an estimate of the probability that, if a given phrase is a key phrase, then a given system will output it as a key phrase.

Corpus	Precision	Recall	f-measure
Corpus 1	53.3	57.6	55.4
Corpus 2	**59.6**	**63.1**	**61.3**
Corpus 3	41.7	48.3	44.8

Table 1 – Results of testing the author's Natural Language ToolKit system

In Table 1, precision, recall, and the f-measure results are shown. The highest precision (59.6), recall (63.1), and f-measure (61.3) were achieved on the smallest sample (19 messages). Since only three sets were evaluated, one cannot determine the coloration between size of the sample and performance of the extractor.

Turney [47] evaluates four key phrase extraction algorithms using 311 email messages collected from 6 employees, and in which 75% of each employee's messages was used for training and 25% (approximately 78 messages) was used for testing. His evaluation approach is similar to the authors of this paper and the highest f-measure reported was that of the NRC, the extractor component of GenEx, which uses supervised learning from examples. The f-measure reported is 22.5, which is, as expected, significantly less than the f-measures shown in Table 1. Hulth [45] reports results from three different term selection approaches. The highest f-measure reported was 33.9 from the n-gram approach with POS tags assigned to the terms as features. All unigrams, bigrams, and trigrams were extracted, after which a stop list was used where all terms beginning or ending with a stopword were removed.

The Natural Language ToolKit system developed by the authors appears to have the best f-measure results in the world when it comes to email knowledge extraction. Although the results are pleasing, the sight of a fully automated system that can extract

knowledge from email without user intervention appears to be many years away, if at all possible. However, with the financial muscle of organisation's like Google developing techniques for their range of information retrieval applications, this domain is likely to see rapid progress within a short period of time.

V. CONCLUSION

This paper has reviewed the four generations of building systems to share knowledge and highlighted the challenges faced by all. The paper discussed the techniques used to extract key phrases and the limitations in the NLP approaches which have defined the boundaries of the domain. The paper has shown that although the f-measure results of the study are encouraging, there is still a requirement for user intervention to enable the system to be accurate enough to provide substantial results to the end users. It is concluded that NLP techniques are still many years away from providing a fully automated knowledge extraction system.

REFERENCES

1. Hiltz, S.R. (1985). Online Communities: A Case Study of the Office of the Future, Ablex Publishing Corp, Norwood, NJ.
2. Lang, K.N., Auld, R. & Lang, T. (1982). "The Goals and Methods of Computer Users", International Journal of Man-Machine Studies, vol. 17, no. 4, pp. 375-399.
3. Mintzberg, H. (1973). The Nature of Managerial Work, Harper & Row, New York.
4. Pelz, D.C. & Andrews, F.M. (1966). Scientists in Organizations: Productive Climates for Research and Development, Wiley, New York.
5. Allen, T. (1977). Managing the Flow of Technology, MIT Press, Cambridge, MA.
6. Cross, R. & Sproull, L. (2004). "More Than an Answer: Information Relationships for Actionable Knowledge", Organization Science, vol. 15, no. 4, pp. 446-462.
7. Kraut, R.E. & Streeter, L.A. (1995). "Coordination in Software Development", Communications of the ACM, vol. 38, no. 3, pp. 69-81.
8. Maltzahn, C. (1995). "Community Help: Discovering Tools and Locating Experts in a Dynamic Environment", CHI '95: Conference Companion on Human Factors in Computing SystemsACM, New York, NY, USA, pp. 260.
9. Campbell, C.S., Maglio, P.P., Cozzi, A. & Dom, B. (2003). "Expertise Identification Using Email Communications", Twelfth International Conference on Information and Knowledge Management New Orleans, LA, pp. 528.
10. Bishop, K. (2000). "Heads or Tales: Can Tacit Knowledge Really be Managed", Proceeding of ALIA Biennial Conference Canberra, pp. 23.
11. Cross, R. & Baird, L. (2000). "Technology is not Enough: Improving Performance by Building Organizational Memory", Sloan Management Review, vol. 41, no. 3, pp. 41-54.
12. Gibson, R. (1997). Rethinking the Future: Rethinking Business, Principles, Competition, Control & Complexity, Leadership, Markets, and the World, Nicholas Brealey, London.
13. Lang, J.C. (2001). "Managing in Knowledge-based Competition", Journal of Organizational Change Management, vol. 14, no. 6, pp. 539-553.
14. Stewart, T.A. (1997). Intellectual Capital: The New Wealth of Organizations, Doubleday, New York, NY, USA.
15. Cross, R. (2000). "More than an Answer: How Seeking Information Through People Facilitates Knowledge Creation and Use", Toronto, Canada.
16. Burt, R.S. (1992). Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge.
17. Erickson, B.H. (1988). "The Relational Basis of Attitudes." in Social Structures: A Network Approach:, Barry Wellman and S. D. Berkowitz (eds.), edn, Cambridge University Press., New York:, pp. 99-121.
18. Schön, D.A. (1993). "Generative Metaphor: A Perspective on Problem-setting in Social Policy" in Metaphor and Thought, ed. A. Ortony, 2nd edn, Cambridge University Press, Cambridge, pp. 137-163.
19. Walsh, J.P. (1995). "Managerial and Organizational Cognition: Notes from a Trip down Memory Lane.", Organizational Science, vol. 6, no. 3, pp. 280-321.
20. Weick, K.E. (1979). The Social Psychology of Organising, 2nd edn, McGraw-Hill, New York.
21. Weick, K.E. (1995). Sense making in Organisations, Sage, London.
22. Blau, P.M. (1986). Exchange and Power in Social Life, Transaction Publishers, New Brunswick, NJ.
23. March, J.G. & Simon, H.A. (1958). Organizations, Wiley, New York.
24. Lave, J. & Wenger, E. (1991). Situated Learning : Legitimate Peripheral Participation, Cambridge University Press, U.K.
25. Yimam-Seid, D. and Kobsa, A. (2003) 'Expert finding systems for organizations: problem and domain analysis and the DEMOIR approach', Journal of Organizational Computing and Electronic Commerce, Vol. 13, No. 1, pp.1–24.
26. Ackerman, M.S. and Malone, T.W. (1990) 'Answer garden: a tool for growing organizational memory', Proceedings of ACM Conference on Office Information Systems, Cambridge, Massachusetts, pp.31–39.
27. Ackerman, M.S. (1994) 'Augmenting the organizational memory: a field study of answer garden', Proceedings of the ACM Conference on Computer-Supported Cooperative Work, pp.243–252.
28. Cohen, A.L., Maglio, P.P. and Barrett, R. (1998) 'The expertise browser: how to leverage distributed organizational knowledge', Presented at Workshop on Collaborative Information Seeking at CSCW'98, Seattle, Washington.
29. Krulwich, B. and Burkey, C. (1996a) 'Learning user information interests through the extraction of semantically significant phrases', In AAAI 1996

Spring Symposium on Machine Learning in Information Access, Stanford, California.
30. Balog, K. and de Rijke, M. (2007) 'Determining expert profiles (with an application to expert finding)', Proceedings of the Twentieth International Joint Conferences on Artificial Intelligence, Hyderabad, India, pp.2657–2662.
31. Maybury, M., D'Amore, R. and House, D. (2002) 'Awareness of organizational expertise', International Journal of Human-Computer Interaction, Vol. 14, No. 2, pp.199–217.
32. Streeter, L.A. and Lochbaum, K.E. (1988) 'An expert/expert-locating system based on automatic representation of semantic structure', Proceedings of the Fourth Conference on Artificial Intelligence Applications, San Diego, California, pp.345–349.
33. Mattox, D., Maybury, M. and Morey, D. (1999) 'Enterprise expert and knowledge discovery', Proceedings of the 8th International Conference on Human-Computer Interaction, Munich, Germany, pp.303–307.
34. Luhn HP. 1958. The automatic creation of literature abstracts. *I.B.M. Journal of Research and Development*, 2 (2), 159-165.
35. Marsh E, Hamburger H, Grishman R. 1984. A production rule system for message summarization. In AAAI-84, Proceedings of the American Association for Artificial Intelligence, pp. 243-246. Cambridge, MA: AAAI Press/MIT Press.
36. Paice CD. 1990. Constructing literature abstracts by computer: Techniques and prospects. Information Processing and Management, 26 (1), 171-186.
37. Paice CD, Jones PA. 1993. The identification of important concepts in highly structured technical papers. SIGIR-93: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 69-78, New York: ACM.
38. Johnson FC, Paice CD, Black WJ, Neal AP. 1993. The application of linguistic processing to automatic abstract generation. Journal of Document and Text Management 1, 215-241.
39. Salton G, Allan J, Buckley C, Singhal A. 1994. Automatic analysis, theme generation, and summarization of machine-readable texts. Science, 264, 1421-1426.
40. Kupiec J, Pedersen J, Chen F. 1995. A trainable document summarizer. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 68-73, New York: ACM.
41. Brandow R, Mitze K, Rau LR. 1995. The automatic condensation of electronic publications by sentence selection. Information Processing and Management, 31 (5), 675-685.
42. Jang DH, Myaeng SH. 1997. Development of a document summarization system for effective information services. RIAO 97 Conference Proceedings: Computer-Assisted Information Searching on Internet; 101-111. Montreal, Canada.
43. Tzoukermann E, Muresan S, Klavans JL. 2001. GIST-IT: Summarizing Email using Linguistic Knowledge and Machine Learning. In Proceeding of the HLT and KM Workshop, EACL/ACL.
44. Tedmori, S., Jackson, T.W. and Bouchlaghem, D. (2006) 'Locating knowledge sources through keyphrase extraction', *Knowledge and Process Management*, Vol. 13, No. 2, pp.100–107.
45. Hulth A. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03). Sapporo.
46. Turney PD. 1997. Extraction of Keyphrases from Text: Evaluation of Four Algorithms, National Research Council, Institute for Information Technology, Technical Report ERB-1051. (NRC #41550)