Inquiry-answer extraction system for disaster risk reduction and the development of sustainable communities

Kaklauskas, A, Amaratunga, RDG and Budryte, L

Title	Inquiry-answer extraction system for disaster risk reduction and the development of sustainable communities
Authors	Kaklauskas, A, Amaratunga, RDG and Budryte, L
Type	Conference or Workshop Item
URL	This version is available at: http://usir.salford.ac.uk/id/eprint/23369/
Published Date	2011

Inquiry-Answer Extraction System for Disaster Risk Reduction and the Development of Sustainable Communities

Arturas Kaklauskas,

Department of Construction Economics and Property Management, Vilnius Gediminas Technical University

arturas.kaklauskas@vgtu.lt

Dilanthi Amaratunga,

Centre for Disaster Resilience, University of Salford

R.D.G.Amaratunga@salford.ac.uk

Loreta Budryte,

Consultus Magnus

lbudryte@gmail.com

Abstract

BELL-CURVE is an EU funded collaborative research project which aims to modernise the Higher Education Institutions (HEIs) in order to make them more responsive to the construction labour market skill needs by continuously improving the skills and knowledge of the construction professionals. To achieve this aim, the Inquiry-Answer Extraction System was developed by using Personalized Multiple Criteria Search Model, which consists from Explicitly and Implicitly Sub-Models. The Personalized Multiple Criteria Search Model stores data that is specific to each individual user. Personalized search depends on a user profile that is unique to the individual. Initial and modification of initial query can be performed by analyzing a system of factors: text materials popularity and reputation, latent semantic indexing, supporting phrases, title and content of the text, behind relevance feedback, keywords density. The Inquiry-Answer Extraction System was adapted for disaster risk reduction and the development of sustainable communities. The Inquiry-Answer Extraction System for Disaster Risk Reduction and the Development of Sustainable Communities is delivering the required quantity of information (amount of pages or number of minutes for reading/reporting/presentation) according to submitted inquiry.

Keywords: Inquiry-Answer Extraction System, Disaster Risk Reduction, Project BELL-CURVE, Case Study.

1. The Inquiry-Answer Extraction System for Disaster Risk Reduction and the Development of Sustainable Communities

BELL-CURVE is an EU funded collaborative research project which aims to modernise the Higher Education Institutions (HEIs) in order to make them more responsive to the construction labour market skill needs by continuously improving the skills and knowledge of the construction professionals. To achieve this aim, the Inquiry-Answer Extraction System was developed. The Inquiry-Answer Extraction System was adapted for disaster risk reduction and the development of sustainable communities. The Inquiry-Answer Extraction System for Disaster Risk Reduction and the Development of Sustainable Communities (Inquiry-Answer Extraction System hereafter) is delivering the required quantity of information (amount of pages or number of minutes for reading/reporting/presentation) according to submitted inquiry.

Electronic knowledge is becoming more and more popular of late, and the information contained in them is constantly enlarging. As the capacity of stored electronic information and knowledge enlarges, it becomes more and more difficult to identify desired materials. In other words, a problem with retrieval effectiveness is encountered. The most advanced electronic libraries of today employ a retrieval system that is similar to Google. The Google system is based on complex PageRank™ technology, which is being patented at this time. It ensures that the most important results always appear first. PageRank objectively assesses the importance of Internet pages. Such a priority is calculated by resolving an equation with 500 million variables and more than 2 billion terms. PageRank utilises a wide Internet structure of reference points as an organisational tool. Essentially Google assesses a reference from Page A to Page B as a "vote" of Page A in behalf of Page B. Google assesses the importance of a page in accordance with the "votes" it has gathered. Google also analyses the page of the "voters". The complicated, automatic Google retrieval method blocks off human interferences. Google is compiled so that no one can purchase a more beneficial position amid the results or otherwise influence the results for commercial purposes.

US Patent No. 5297042 discloses document retrieval system that includes an inputting unit for inputting a retrieval condition including one or a plurality of keywords and a weight value for each keyword, an operating unit having first factors corresponding to relationship values, each relationship value being defined as a degree of the relationship between two keywords out of keywords which are predetermined in the document retrieval system and second factors corresponding to importance values, each importance value being defined as a degree of importance of a keyword in each one of a plurality of documents wich are predetermined in the document retrieval system, the operation unit generating a relevance value, which represents a degree of relevance in satisfying a user's requirement, for each of the documents on the basis of the retrieval condition input from the inputting unit, the first factors and the second factors, and an outputting unit for outputting the relevance value fore each of documents as a retrieval results.

US Patent Application US 2006/0047656 A1 discloses a computer-readable code, system and method for retrieving one or more selected text materials from a library of documents. The system processes a user-input search query representing the content of the text to be retrieved, and accesses a word index for the documents to identify those text materials in the database having the highest word-match scores with the search query. The weights of words in the query may be adjusted to optimize the search.

European patent EP 1880320 (WO 2006121338) discloses a method for determining contextual summary information across documents in a search result comprising steps for a) applying a query to a set of documents, b) identifying matching documents, c) selecting for the matching documents a query-dependent subsection (chapters/sections/paragraphs) of each of document, d) selecting document properties associated with the document subsection; and computing summary information for the document properties across a result set (performing multi-criteria analysis of selected composite parts according to the initial query).

The known answer extraction and retrieval systems does not perform design of the set of alternative variants of summarised text by permutating each composite part with every other composite part nor selects out the best alternative variant from the alternative variants of summarised text developed from different text material from a library of documents according to the initial query or according to the modified initial query performed by request of actual user. Thus the known answer extraction and retrieval systems are not enough flexible in search and not enough informative in selecting out and integrating rational electronic information.

The developed Inquiry-Answer Extraction System has been made in view of the above problems, and it is an object of the present development to provide an electronic information answer extraction and retrieval system that would be more flexible and more informative in selecting out and integrating rational electronic information, as much by the desired area as by coverage, and that would allow the actual user to participate and have an influence during the time of the operations by automatically designing, evaluating and selecting the information most suitable for him/herself.

Search engine rankings have been adopted in most advanced intelligent libraries (Alexandrov *et al.*, 2003, Gutwin *et al.*, 1999, Hsinchun *et al.*, 1998, Kaklauskas *et al.*, 2006, Ruch *et al.*, 2007, Trnkoczy *et al.*, 2006, J. Wang, 2003) and tutoring systems (Armani *et al.*, 2000, Brusilovsky, 2000, Day *et al.*, 2007, Lucence, 2005, and Pouliquen *et al.*, 2005). As part of the ongoing Illinois Digital Library Initiative project, research proposes an intelligent personal spider (agent) approach to Internet searching, which is grounded on automatic textual analysis, general-purpose searches and genetic algorithms (Hsinchun *et al.*, 1998). Pouliquen *et al.* (2005) that uses parsing techniques to extract information from texts, and provide a proper semantic indexation that is used by a medical-specific search engine. Day *et al.* (2007) use the Jakarta Lucene full-text indexer to index full-texts of textbooks. Jakarta Lucene is a high-performance, fully-featured text search engine library written entirely in Java. Its technology is suitable for nearly all applications that require full-text searches. ITA (Pouliquen *et al.*, 2005) index chapters, sections, and subsections of textbooks. Highlighters are used to highlight the

index context. Finally, the ITA provides reading recommendations for students via a chapter similarity function. However, intelligent libraries (Alexandrov *et al.*, 2003, Gutwin *et al.*, 1999, Hsinchun *et al.*, 1998, Kaklauskas *et al.*, 2006, Ruch *et al.*, 2007, Trnkoczy *et al.*, 2006, J. Wang, 2003) and intelligent tutoring systems (Armani *et al.*, 2000, Brusilovsky, 2000, Day *et al.*, 2007, Lucence, 2005, and Pouliquen *et al.*, 2005) with search engine rankings cannot select chapters (sections, paragraphs) of specific texts, which are the most relevant to a student, cannot integrate the chapters (sections, paragraphs) of specific texts into learner specific alternatives of teaching material and cannot select the most rational alternative, i.e. cannot develop alternatives of training materials, perform multiple criteria analysis and automatically select the most effective variant. However, an Inquiry-Answer Extraction System can perform the afore-mentioned functions. To the best of our knowledge the above function has not been implemented before, and so this attempt is the first time someone has done so. The proposed approach helps students to obtain suitably tailored material for any e-learning course. The above-mentioned and other improvements are possible when using the Inquiry-Answer Extraction System.

Indexing often is used to refer to the automatic selection and compilation of 'meaningful' keywords from e-books into a list that can be used by a search system to retrieve texts. This list is more properly called a concordance. As this procedure involves no intellectual effort indexers distinguish their own work by calling it intellectual indexing, manual indexing, human indexing or back-of-book-style indexing. Indexing also means the intellectual analysis of the e-books to identify the concepts represented in the document and the allocation of descriptors to allow these concepts to be retrieved. During indexing the Inquiry-Answer Extraction System visit definite information/knowledge and collects information/keywords about it. Intelligent copy and paste from many modules with retention of a link/reference to the module can be performed. Development of an answer is performed by using a combination of knowledge found with the possibility of easy editing and integrating. Learners can use Inquiry-Answer Extraction System for computer-assisted extraction of data from text for their own purposes, making their work more efficient. As importantly, these data can then be reused and made useful for a large learners community: they can be incorporated (connected, interlinked) into a large distributed knowledge base.

Different electronic information and knowledge were used during testing of the Inquiry-Answer Extraction System as follows:

1. SUSTAINABLE COMMUNITIES DEVELOPMENT:
    1.1. Livelihoods Restoration
    1.2. Healthy Environment
    1.3. Housing, Environmental Reconstruction
        1.3.1. Coastal Zoning
        1.3.2. Land-use planning
        1.3.3. Reconstruction standards
        1.3.4. Temporary Housing

2. Personalized Multiple Criteria Search Model

The Inquiry-Answer Extraction System was developed by using *Personalized Multiple Criteria Search Model*, which consists from Explicitly and Implicitly Sub-Models. The *Personalized Multiple Criteria Search Model* stores data that is specific to each individual user. Personalized search depends on a user profile that is unique to the individual. *Explicitly Sub-Model (Sub-System)* learns by asking the user. *Implicitly Sub-Model (Sub-System)* learns by observing the user's historical query. The *Implicitly Sub-Model (Sub-System)* starts by assessing the user's required subject information, individual's search and web history and content the user has read in order to adjust its search results based on past behaviour. Initial and modification of initial query can be performed by analyzing a system of factors: text materials popularity and reputation, latent semantic indexing, supporting phrases, title and content of the text, behind relevance feedback, keywords density.

Relevance feedback refers to methods for adjusting a search statement based on preliminary relevance judgments by the user. The usual approach is for a preliminary search to proceed using terms (and modifications such as term weights, truncation, proximity limits, etc.) provided by the user. The results of this initial search are presented to the user, along with an evaluative questionnaire in which the user can indicate preliminary relevance judgments concerning the value of the retrieved documents. These judgments are then used by the system to modify the initial search statement (e.g., adding weights to the more successful terms, decreasing weights for the less successful terms or eliminating them altogether), and a second search is performed. This interaction can continue as long as the user wishes.

Personalized Multiple Criteria Search Model feedback provides a method for e-intelligent analysis. It automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction. Personalized Multiple Criteria Search Model feedback automates the manual part of relevance feedback and has the advantage that assessors are not required. Through a query expansion, some relevant documents missed in the

initial round can then be retrieved to improve the overall performance. Clearly, the effect of this Model strongly relies on the quality of selected expansion terms. In addition, if the words added to the original query are unrelated to the query topic, the quality of the retrieval is likely to be degraded. Specifically, the proposed Model assigns more weights to words occurring closer to query words based on the intuition that words closer to query words are more likely to be related to the query topic.

The Model involves a combination of On-page factors (such as keyword frequency, headings, links, etc.) and Off-page factors (hyperlink analysis, etc.) factors which produce a distinct effect (a higher ranking for specific exact match and broad match shingles).

The article details a Case Study of the testing of the developed Inquiry-Answer Extraction System as follows.

3. Case Study

The Department of Construction Economics and Property Management of Vilnius Gediminas Technical University introduced the following programs in e-learning for Master's degree studies: Real Estate Management in 1999, Construction Economics in 2000, Internet Technologies and Real Estate Business in 2003 and Intelligent Built Environment in 2009. There are currently 239 master students from all over Lithuania studying in these four e-learning master's programs. The initial version of the Inquiry-Answer Extraction System was developed in 2007. The testing of the System has been ongoing since that time. There were 26 e-learning students testing the System. The System was improved in light of the continuous testing results.

A case of an e-learning student from the Department of Construction Economics and Property Management of Vilnius Gediminas Technical University is analysed to illustrate operation of the developed System. Initially student open the website http://iti.vgtu.lt/alib/default.aspx to access the Inquiry-Answer Extraction System.

To help the system choose the most appropriate text from numerous books available in the System's library, search keywords must be used. There are two ways to enter search keywords:

- The search keywords available in the database: typical search keywords are listed in the field on the right. Left-click to mark the search keyword of your choice and enter its weight, if any, in the field *Weight for keywords adding* (if no weight is specified, the system uses the default value 1). Then click the button *Add selected keyword from the list*. Repeat these actions until all search keywords of your choice are listed in the top left field. If the search keywords have synonyms, they are also shown among the search keywords in the top left field.
- New search keywords. Type your search keyword in the field *New keyword* and specify its weight, if any, in the field *Weight*. Then click the button *Add a keyword*. Repeat these actions until all search keywords of your choice are listed in the top left field.
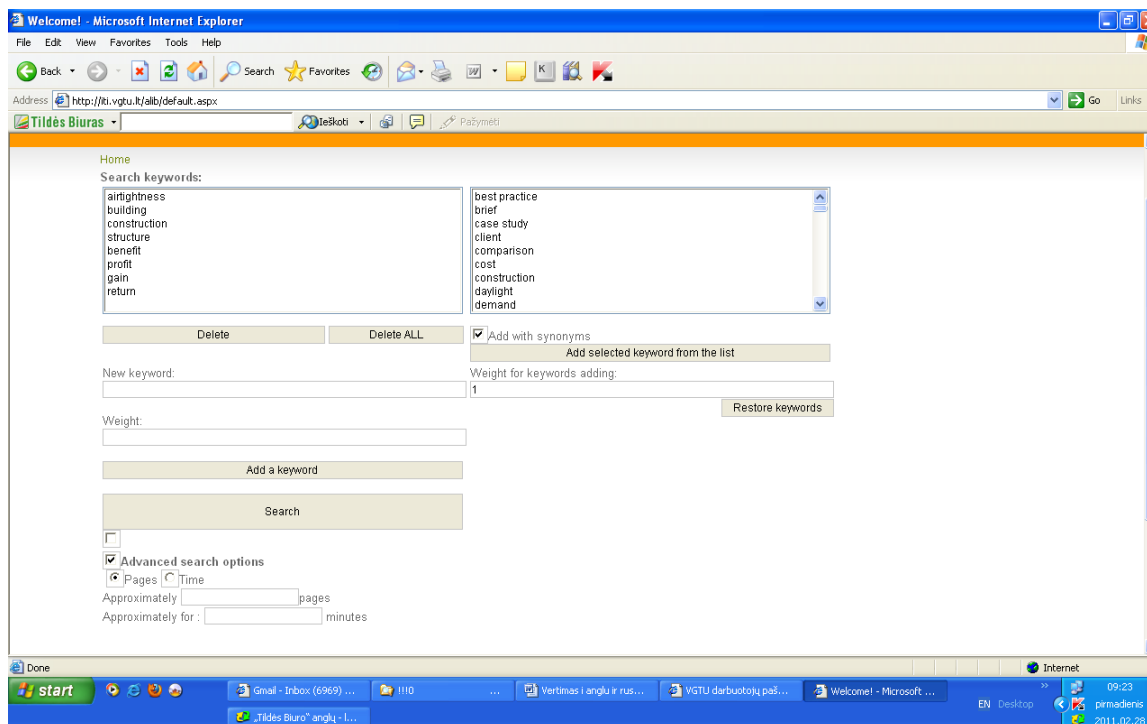
Figure 1: The Inquiry-Answer Extraction System

Student can search for the most appropriate text from a range of books by using three alternatives:

- The search restricted by the number of pages. Tick the box ☑ next to *Advanced search options* and the System will offer to restrict the search by the number or pages or minutes. Enter the number of pages of your choice in the field *Approximately* [ ] *pages*, then click Search. The System will show the requested number of pages.
- The search restricted by the number of minutes (individual reading, reading to students). Tick the box ☑ next to *Advanced search options*, then enter the number of minutes of your choice in the field *Approximately for:* [ ] *minutes*. The System will show the requested amount of text.
- A click on the *Search* button launches the search and the system combines all possible alternatives and displays the top 20 variants. The system lists the alternatives by their priority and compiles them from the paragraphs taken from three books available in the System's database and most adequate to the user's needs.

Table 1 shows the frequency of each specific keyword in the analysed text. Keyword ranking in modules seeks to determine the level of relevance of chapters and sections for student's needs. The level of relevance to student's needs can be defined by the term "Keyword density and significance" as described by indicators provided in the table: weight (shows the significance of one keyword over another from a student's perspective in a search for specific learning material), difficulty of a text (the level of difficulty is determined on the basis of previous examination results related to a specific topic) and other indicators (number of pages, words and sentences in the analysed text) which help to determine the keyword's density. Then,

information describing the usefulness of the analysed text for a learner's needs is summarised in a Table 1. Also, the relevance of a text to a student's learning needs is described by the presence of different keywords in one sentence (see Table 2). The occurrence of several different keywords that are specified by a student in the same sentence shows higher relevance of the text to the learner's needs.

The significance/efficiency (Q_j) of alternatives of the teaching material is determined on the basis of keyword density characteristics (i.e. frequency of each specific keyword, weight, difficulty of a text, number of pages, words and sentences in the analysed text). Significance Q_j of the learning material aj indicates the satisfaction degree of requirements and goals pursued by the students, e.g. the greater the Q_j the higher the efficiency of the learning material.

Table 1: Density of specific keywords in the analysed text

Legend - The first column of the table provides the keywords under evaluation; the second column provides the weight of the keywords. The table's third column specifies whether the minimising or maximising value is the best. From the fourth column onwards, numbers of paragraphs are provided with references to full texts and the frequency of iterated keywords.

Table 2: Combinations of analysed keywords in sentences

Repetitions	A human rights assessment 1	A human rights assessment 2	A human rights assessment 3	A human rights assessment 4	A human rights assessment 5	A human rights assessment 6
1*	6	12	9	28	29	21
2*	9	11	5	8	10	1
3*	2	2	3	3	6	2
4*	3	0	1	0	0	0
5*	1	0	0	0	0	0
6*	0	0	0	0	0	0
7*	0	0	0	0	0	0
8*	0	0	0	0	0	0
9*	0	0	0	0	0	0

Legend - *The first column of the table provides the number of iterations in one sentence. From the second column onwards, numbers of paragraphs are provided with references to full texts and the number of iterated different keywords in one sentence.*
1* - numbers in the row show the frequency of each specific keyword in the analysed text;
2* - numbers show the frequency of combinations of two different keywords in one sentence in the analysed text;
3* - numbers show the frequency of combinations of three different keywords in one sentence in the analysed text, etc.

The degree of utility N_j of the teaching material aj indicates the level of satisfying the needs of the actual student. The more learning goals that are achieved and the more important they are, the higher the degree of the teaching material's utility. The degree of the teaching material's utility reflects the extent to which the goals pursued by the student are attained. The greater the Q_j the higher the priority of the teaching material (see Table 3).

Table 3: Results of Multiple Criteria Evaluation

Inquiry-Answer Extraction System can display previously covered keywords that might be used for search required knowledge. The tutor can add additional keywords to this list. Also the search is possible by any combination of keywords. Using the keywords provided by a student the system formulates a number of alternatives for a required teaching material. These alternatives are composed from sections or components of many different modules matched in a certain way. The selection of keywords and determination of their importance is not as simple as it seems. Numbers of feasible alternatives can be as large as 100 000 000.

Conclusions

During BELL-CURVE project the Inquiry-Answer Extraction System was developed by using Personalized Multiple Criteria Search Model, which consists from Explicitly and Implicitly Sub-Models. The Personalized Multiple Criteria Search Model stores data that is specific to each individual user. Personalized search depends on a user profile that is unique to the individual. The System is able to search and find useful material, carry out a multivariant text design, multiple criteria analysis and select the most rational text material alternatives according to individual users' demands. Initial and modification of initial query can be performed by analyzing a system of factors: text materials popularity and reputation, latent semantic indexing, supporting phrases, title and content of the text, behind relevance feedback, keywords density.

References

Alexandrov V N, Dimov I T, Karaivanova A and Tan C J K (2003) "Parallel Monte Carlo algorithms for Information Retrieval." *Mathematics and Computers in Simulation* **62**(3-6): 289–295.

Armani B, Bertino E, Catania B, Laradi D, Marin B and Zarri G P (2000) *Repository Management in an Intelligent Indexing Approach for Multimedia Digital Libraries*, In: Ohsuga, S., Raś, Z.W. (eds.) ISMIS 2000, LNCS (LNAI), vol. 1932, Springer, Heidelberg.

Brusilovsky P (2000) *Course Sequencing for Static Courses? Applying ITS Techniques in Large-Scale Web-Based Education*, In: Gauthier, G., VanLehn, K., Frasson, C. (eds.) ITS 2000, LNCS, vol. 1839, pp. 19–23. Springer, Heidelberg.

Day M, Lu C, Yang J, Chiou G, Ong C and Hsu W *Designing an Ontology-Based Intelligent Tutoring Agent with Instant Messaging* (available online http://iasl.iis.sinica.edu.tw/webpdf/paper-2005-Designing_an_Ontologybased_Intelligent_Tutoring_Agent_with_Instant_Messaging.pdf [accessed on 28/01/2009])

Gutwin C, Paynter G, Witten I, Neville M C and Frank E (1999) "Improving Browsing in Digital Libraries with Key-phrase Indexes." *Decision Support Systems* **27**(1-2): 81–104.

Hsinchun C, Yi-Ming C, Ramsey M and Yang C C (1998) "An Intelligent Personal Spider (Agent) for Dynamic Internet/Intranet Searching." *Decision Support Systems* **23**(1): 41–58.

Kaklauskas A, Zavadskas E and Ditkevicius R (2006) *An Intelligent Tutoring System for Construction and Real Estate*, In: Luo, Y. (ed.) CDVE 2006, LNCS, vol. 4101, pp. 174–181. Springer, Heidelberg.

Lucence J *Jakarta Lucene Text Search Engine in Java* (available online http://jakarta.apache.org/lucene/docs/index.html [accessed on 23/02/2009]).

Pouliquen B, Le Duff F, Delamarre D, Cuggia M, Mougin F and Le Beux P (2005) "Managing Educational Resource in Medicine: System Design and Integration." *International Journal of Medical Informatics* **74**(2-4): 201–207.

Ruch P, Boyer C, Chichester C, Tbahriti I, Geissbühler A, Fabry P, Gobeill J, Pillet V, Rebholz-Schuhmann D and Ovis C (2007) "Using Argumentation to Extract Key Sentences from Biomedical Abstracts." *International Journal of Medical Informatics* **76**(2-3): 195–200.

Trnkoczy J, Turk Z and Stankovski V (2006) "A Grid-Based Architecture for Personalized Federation of Digital Libraries." *Library Collections, Acquisitions, and Technical Services* **30**(3-4): 139–153.

Wang J (2003) "A Knowledge Network Constructed by Integrating Classification, Thesaurus, and Metadata in Digital Library." *The International Information & Library Review* **35**(2-4), 383–397.