Communications of the Association for Information Systems

Volume 30 Article 26

6-2012

Examining Question-Answering Technology from the Task Technology Fit Perspective

José Antonio Robles-Flores
Graduate School of Business, ESAN University, jrobles@esan.edu.pe

Dmitri Roussinov
Department of Computer and Information Sciences, University of Strathclyde

Follow this and additional works at: https://aisel.aisnet.org/cais

Recommended Citation
Robles-Flores, José Antonio and Roussinov, Dmitri (2012) "Examining Question-Answering Technology from the Task Technology Fit Perspective," *Communications of the Association for Information Systems*: Vol. 30 , Article 26.
DOI: 10.17705/1CAIS.03026
Available at: https://aisel.aisnet.org/cais/vol30/iss1/26

This material is brought to you by the AIS Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in Communications of the Association for Information Systems by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Communications of the Association for Information Systems

CAIS

Examining Question-Answering Technology from the Task Technology Fit Perspective

José Antonio Robles-Flores

Graduate School of Business, ESAN University

jrobles@esan.edu.pe

Dmitri Roussinov

Department of Computer and Information Sciences, University of Strathclyde

Abstract:

The World Wide Web has become a vital supplier of information for organizations in order to carry on such tasks as business intelligence, security monitoring, and risk assessments. By utilizing the task-technology fit (TTF) theory, we investigate the issue of when open-domain question-answering (QA) technology would potentially be superior to general-purpose Web search engines. Specifically, we argue theoretically and back up our arguments with a user study that the presence of *fusion* (information synthesis) is crucial to warrant the use of QA. At the same time, many information seeking tasks do not require fusion and, thus, are adequately served by traditional keyword search portals (Google, MSN, Yahoo, etc.). This explains why prior attempts to demonstrate the value of QA empirically were unsuccessful. We also discuss methodological challenges to any empirical investigation of QA and present several solutions to those challenges, validated with our user study. In order to carry our study, we created a novel prototype by following the Design Science guidelines. Our prototype is the first of its kind and is capable of answering list questions, such as *What companies own low orbit satellites*? or *In which cities have illegal methyl-methionine labs been found?* This investigation is only a precursor to a full-scale empirical study, but it serves as a medium to overview the state of the art QA technologies and to introduce important theoretical and empirical concepts involved. Although we did not find empirical evidence that one technology is uniformly better than the other, we discovered that once the user accumulates experience using QA, he/she can make an intelligent decision whether to use it for a particular task, which leads to the user to be more productive on average with the same tasks compared to when there is no choice of technology.

Keywords: design science, decision support systems, exploratory research, task technology fit, question answering, business intelligence, World Wide Web, Internet search engines, natural language processing

Volume 30, Article 26, pp. 439-454, June 2012

The manuscript was received 8/17/2010 and was with the authors 6 months for 1 revision.

I. INTRODUCTION

A critical component for the success of a modern enterprise is its ability to take advantage of all the available information in a timely manner. The World Wide Web represents a rich source of information and allows access to knowledge that was not readily available before [Katz et al., 2004; Lyman and Varian, 2000; Roussinov et al., 2008]. Organizations benefit from the Web as a source of information in different ways: it contains information about customers' perception of the market, the history and reputation of potential competitors or suppliers, and, in today's increasingly global economy, the background about the countries and cultures where a company may choose to operate. Technical personnel typically use the Web to look for the tips and ideas on solving common IT and more general business problems or for reviews of important trends [Robles-Flores, 2009].

Since the average computer user spends half an hour each day searching the Web by using the popular portals, it is not surprising that the leading portals (e.g., Google and Yahoo together) started to rival the prime-time advertising revenues of America's three big television networks: ABC, CBS, and NBC [Mills, 2005]. The three most popular search portals (Google, Yahoo, Microsoft's Bing) are all among the top ten most popular sites [Alexa, 2010]. Their success prompts investors and IT practitioners to ask the question, "What comes next?". Although it is hard to provide a definite answer to this at the moment, we can find at least a hint in the widely publicized Jeopardy match between the IBM computer called Watson (currently the fastest, as well as having the largest memory in the world) and the best human players [CNN Money, 2011].

Web search engines are commonly used to locate information for business analysis; however, they typically retrieve a large number of pages only to overload business analysts with irrelevant information [Chung et al., 2005; Lyman and Varian, 2000; Roussinov et al., 2008]. More fine-grained technologies capable of understanding (for example, Business Intelligence (BI) tasks) and representing their results in comprehensible format are emerging. Among these emerging technologies is the automated Question Answering (QA) technology, which serves to locate, extract, and represent a specific answer to a user question expressed in natural language. A QA system takes an input, such as "How many cars are sold in Turkey?" and provides an output, such as "2,000 to 3,000 vehicles are sold in Turkey each year," or simply "2500."

In spite of breakthroughs, a fully automated question-answering (QA) system remains an extremely challenging task. On the other side, keyword-based search portals, such as Google, Yahoo, and MSN, have significantly improved their accuracy at pointing right pages by counting the number of occurrences of the query words, estimating popularity of the pages on the Web [Brin and Page, 1998] and filtering those created by spammers. The search engines, however, are still not designed to handle questions. Instead, their algorithms are based on the classical "bag of words" model, which ignores the order of keywords. For example, the query "Who is the largest producer of software?" will be treated exactly the same as "largest software producer," which brings non-intuitive results: pages about largest producers of dairy products, trucks, and "catholic software," but not the answer you would expect (e.g., "Microsoft"). Thus, even if the correct answer is among the search results, the user still needs to review the output "snippets" to locate it.

Most of the research on automated question-answering systems has been focused on the underlying algorithms. The results from TREC [Voorhees and Buckland, 2007] annual evaluations—the standard benchmarks in the community of QA researchers—have demonstrated that the state of the art QA systems are indeed capable of delivering accuracy superior to keyword-based information retrieval systems. However, the tests so far have been performed in a "batch" mode only, leaving the interaction between the user and the system completely out of the picture. The outcomes of the empirical tests involving users have so far been inconclusive [Voorhees and Buckland, 2007].

As a result, *it is still not clear, whether QA actually helps and what its impact is on several types of business tasks (e.g., business intelligence tasks).* Indeed, since it is often possible to send a question or handpicked keywords to a search engine (e.g., Google) and "eye-ball" the correct answer, many researchers and practitioners doubt whether QA is ever going to make an impact large enough to become a "killer application" and the "Web search of tomorrow" as it was predicted [Clayburn, 2005]. More formal evaluations [Roussinov et al., 2008] show that a correct answer to a question submitted verbatim as a query indeed frequently occurs within the top snippets returned by a Web search engine. The QA promise and the need for its empirical evaluation have been articulated by other researchers [Maybury, 2003], but not yet carried out in practice with the exception of narrow domains: e.g., Roussinov and

Robles-Flores [2007] established that QA can be more effective in locating malevolent Web content than a regular search engine.

As we elaborate further in this article, we suggest two possible explanations for the lack of successful empirical investigation of the QA technology: (1) The top-of-the-line systems are not available for public use; thus the studies involving them are hard to conduct and replicate. (2) The theoretical framework to perform such a study is still nonexistent. This article shows that those two challenges can be overcome. Specifically, we claim the following contributions:

1. We show how the Task-Technology Fit (TTF) model [Goodhue and Thompson, 1995] can serve as the theoretical framework for evaluating QA and its applicability for the business intelligence (and similar) tasks.
2. By applying TTF, we explain why so far QA has not been shown to be more effective than traditional keyword searching (KW). We formalize the notion of *information fusion* and argue that its presence within the search task is necessary for QA to add value. Previous research ignored this crucial component, which can explain the lack of successful or conclusive findings.
3. By following the guidelines of Design Science Research [Hevner et al., 2004], we have implemented and tested a research prototype, the first to our knowledge Web question-answering system capable of handling list questions, for example, *What companies own low orbit satellites?* or *In which cities have illegal methyl-methionine labs been found?* We define *list questions* as questions that expect the answers to be stated in the form of a list of items. This is consistent with the related literature (e.g., Maybury, 2003). This definition also excludes *why* or *how to* questions, since they require different treatment by the current algorithms and are not supported in the system involved here. The prototype is based on the replicable technology and publicly open (not proprietary) algorithms.
4. To validate our evaluation methodology, we have performed a large-scale user study with our prototype. The users were able to compare searching for answers using our prototype and the traditional keyword-based search engine (Google) on a set of information-seeking tasks. These tasks were also suggested by users. We report several important observations and lessons learned from our study.

The next section presents the Task-Technology Fit (TTF) theory as our theoretical framework, followed by a section describing our prototype, followed, in turn, by the section describing our user study. The last section presents conclusions, limitations, and possible future directions.

II. THEORETICAL FRAMEWORK

Task-Technology Fit

The Task-Technology Fit (TTF) model was first introduced into the Information Systems (IS) literature to help understand the link between information systems and individual performance. Goodhue and Thompson [1995] presented and tested the Task-Technology Fit model that focuses on measuring how a certain system fits certain tasks. The fit was defined by the correspondence (matching) between the capabilities of the technology and the requirements of the task: "the degree to which a technology assists an individual in performing his or her portfolio of tasks." Goodhue et al. [1995] established that this "fit" had a direct impact on individual performance and utilization of the system. The latter, in turn, also impacted the performance. Eight components of the fit were proposed and measured by Goodhue et al. [1995] through a questionnaire. An instrument was developed later based on decision-making tasks to measure the fit in twelve dimensions [Goodhue, 1998]. The fit was assessed by experts in Zigurs and Buckland [1998]. In the follow-up work [Zigurs et al., 1999], a categorization of tasks was used along with characteristics of the technology.

A wide range of tasks has been used to evaluate TTF in the studies that followed, including software maintenance [Dishaw and Strong, 1998a; Dishaw and Strong, 1998b; Dishaw and Strong, 1999], answering managerial questions [Goodhue et al., 2000], decision making and daily transaction tasks [Goodhue, 1998; Goodhue and Thompson, 1995], managerial tasks using quantitative information [Goodhue 1995], tasks related to searching a library catalogue [Staples and Seddon, 2004], information seeking [Chen et al., 2006], tasks related to academic research [Goodhue et al., 1997], online shopping [Klopping et al., 2004], and group support systems [Zigurs and Buckland, 1998; Zigurs et al., 1999]. However, *none of the tasks involved finding answers on the Web nor aggregation of information from multiple sources (fusion),* which is the focus of this work.

Information Seeking and Fusion

Information seeking has been defined as a cognitive process to acquire information [Marchionini, 1995]. Characteristics associated with information seeking include: (1) A high level process "in which humans purposefully engage in order to change their state of knowledge" [Marchionini, 1995], (2) Purposive behavior [Wilson, 2000], and (3) "Conscious effort to acquire information in response to a need or gap" [Case, 2002].

Information-seeking tasks often have a specific item of information as the target, for example the address of *The Sunday Times* newspaper in London. Sometimes, however, people may need to compile a list of all newspapers in London. We define the "Fusion of Information" as the *use of multiple sources in order to identify a complete set of the items needed*, for example to answer a specific question. Although the idea of combining information from different sources is not new, our current study is the first in considering it an important dimension in the context of Task-Technology Fit.

Keyword Searching

Today's most common technology to find information on the Web is keyword searching (KW), supported by popular information seeking portals such as Google, Yahoo, and Microsoft's Bing. This technology is freely available and enjoys growing popularity among Internet users [Alexa, 2010]. Although it has evolved to the point of satisfying most of the users, the need for its further improving is also often noted.

We consider two perspectives to look at today's Web search engines: (1) from the *user's perspective*, KW is characterized by the use of keywords as input, e.g., "Denver Aspen," and the output in the form of a list of snippets (brief paragraphs/sentences, extracted from Web documents), which include the links to the actual pages, (2) from the *algorithm perspective*, its objective is to match the input keywords with words in Web documents, taking into consideration the link structure and the popularity of the pages implied by that structure [Brin and Page, 1998].

Question Answering

A new alternative to find information on the Web is automated question answering (QA). This research focuses on completely automated, open (not restricted) domain, Web-based question answering. As it typically is in the literature on question answering [Maybury, 2003], "open domain" means that the question topics are not restricted, as, for example, they can be to medical, biological, or other domains. Thus, in open domain QA, anything can be asked and any source can be used to provide an answer.

The goal of a question-answering system is to retrieve answers to questions, rather than full documents or passages, as most information retrieval systems currently do [Dumais et al., 2002]. Question Answering (QA) technology locates, extracts, and presents a specific answer to a user question expressed in natural language. For example, a QA system takes as input "How big is our galaxy in diameter?" and produces the output "Our galaxy is 100,000 light years in diameter" along with link(s) to source page(s).

Web Question Answering (WebQA) technology is a technology that uses the entire Web as the source of answers. Fully automated Web question-answering systems are still under development, in spite of occasional claims of success by the popular search portals.

As we did with keyword searching technology, in order to study the differentiating characteristics of the Web QA technology, we consider two perspectives as well: (1) From the *user's perspective*, it takes as input a question in a natural language and produces output consisting of snippets (short paragraphs). (2) From the *perspective of the algorithm*, Web QA technology uses all the words entered in the question. Common question patterns are recognized and used to find the answers matching previously learned answer patterns.

An important type of QA system is the one that can support list questions as defined in the Introduction section. Unless the complete set of answers is available from a single source (page), supporting list questions requires fusion of information as defined above.

Comparing: The Fusion and the Fit

In this section, we argue that the presence of fusion is crucial for the QA technology to have a better fit than KW technology. An earlier study on TTF [Goodhue, 1995] confirmed the relevance of the "fit" construct to assess the value of an information system for supporting decision making tasks. A later instrument was developed [Goodhue, 1998] to measure fit along twelve dimensions as a refinement of the earlier eight-factor version. Zigurs and Buckland [1998] defined fit in terms of profiles composed of consistent task contingencies (outcome multiplicity, solution scheme multiplicity, conflicting interdependence, and solution scheme/outcome uncertainty), and system elements.

In general, TTF research describes two ways to assess the fit [Staples and Seddon, 2004]: (a) facets-of-fit approach, and (b) predicted-outcomes approach. Our research follows the former. An example for the facets-of-fit approach is the task of cooking spaghetti with two possible sets of tools [Staples and Seddon, 2004]. Toolset One is a large metal pot and a gas cooker, while Toolset Two is a plastic bowl and open fire. The facets of the task requirements are (a) the container should hold sufficient water, (b) there should be a reliable, controllable heat source, and (c) the container should withstand the heat for ten to fifteen minutes. Toolset One meets all three requirements, while Toolset Two clearly fails the third requirement, as it will melt with the heat, and may fail the second requirement if the open fire cannot be controlled.

Table 1: Summary of Differentiating Characteristics of Technologies for Finding Answers on the Web: KW vs. QA

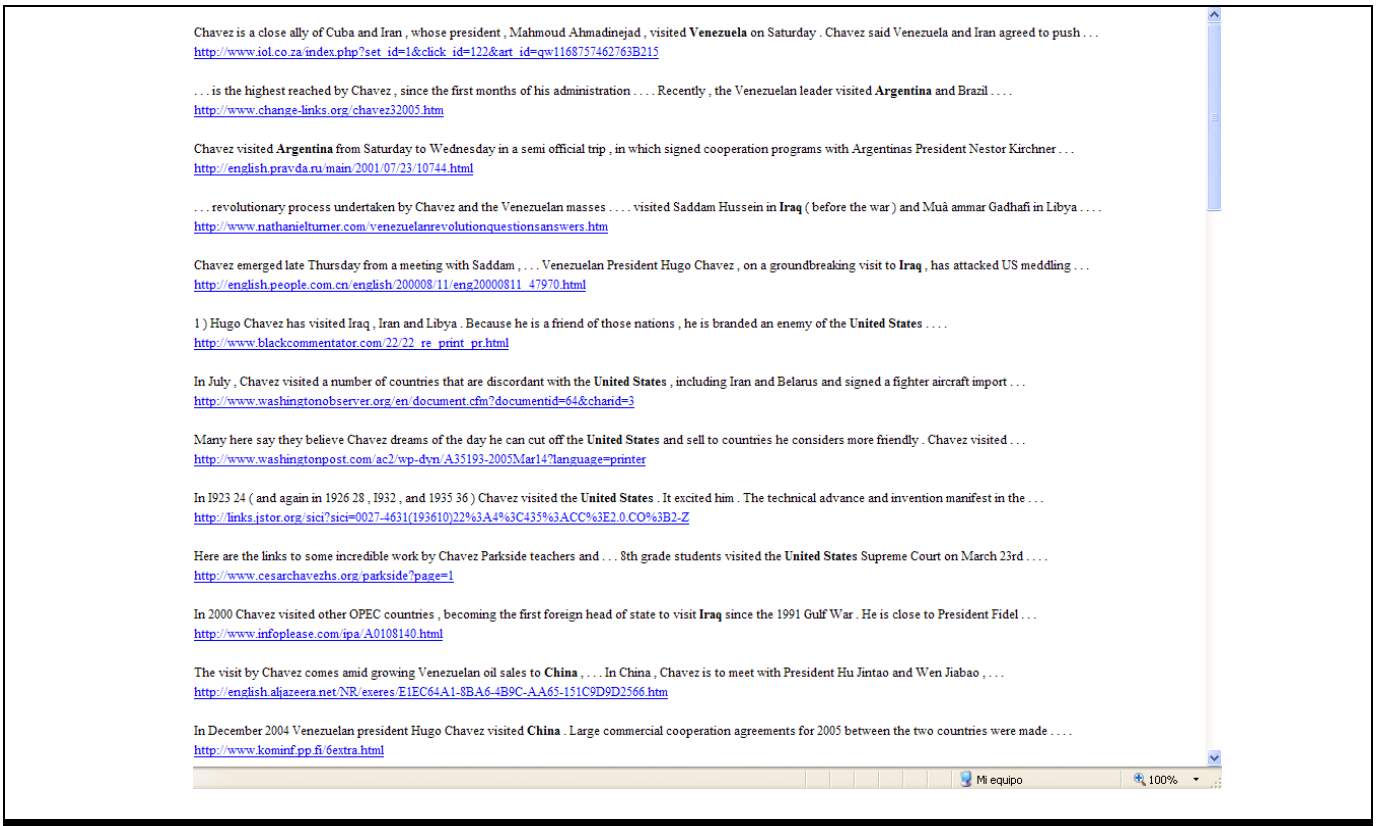
Facets	Web keyword searching	Web question answering
Input format	Keywords	Question in natural language
Output format	Snippets with keywords highlighted, links to the pages	Snippets with the presumed answers highlighted, links to the pages
Ordering of items	By algorithm's perception of the relevance	By algorithm's perception of the value to the user: avoiding redundancy and promoting more convincing snippets
Treatment of input	Bag of words: retrieved documents have to match the keywords	Natural language question: Retrieved documents have to provide answers
Combining results from different sources	Can happen only accidentally if there is high variability in the snippets	Intentionally supported. The snippets are enforced to be non-redundant to each other.

For a task that requires building a list of items to answer a question, the applicable facets from the prior literature are the following: (1) The tool should find items related to the question. (2) The tool should present as many correct (excluding inaccurate and outdated) answers as possible. (3) The tool should present the items in a way that is effective to the user, which typically means the answers have to be (a) recognizable in the output (b) non-redundant (c) convincing. Those requirements are supported by prior empirical studies with QA, e.g., by Roussinov et. al. [2008].

Table 1 summarizes the main differences between the two technologies. In spite of those differentiating characteristics, it is evident that *neither is universally superior to the other when no fusion is involved* for the following reasons*:* (1) both tools can find the answers (QA explicitly by design while KW implicitly by frequent co-occurrence of the answers with the questions words); (2) since a single source may exist for all the answers, there is no strong need to reduce redundancy and to promote diversity of answers; and (3) the perception of the input and output format is the same by the user.

The situation changes significantly when the need for fusion is present as it is the case with list questions. When the tool does not provide the fusion of information, the user needs to manually search multiple sources and put together the disparate pieces. For example, compare the two outputs presented in Figures 1 and 2 showing the system responses to the question *Which countries has Hugo Chavez visited*? Since KW simply looks for the popular pages that frequently mention the words *countries, visited, Hugo, Chavez*, it, not surprisingly, chooses those that mention *Venezuela*, which happens to be wrong. Only one correct answer (*Cuba*) is mentioned by chance. On the contrary, since QA promotes diverse answers (by the algorithms detailed in Section 4), it correctly reports several different countries, including *Cuba, Iran, Argentina, United States,* and *China*. Thus, while KW mostly repeats the same (most popular) answers, QA *avoids redundancy* by noting what answers have been already presented so a more diverse set of answers is reported. In order to recognize which answers have been already reported, a system would need to identify the set of candidate answers and to perform *triangulation* (confirmation), the two important steps not present in KW but intrinsic parts of QA. Moreover, the triangulation in the case of fusion cannot be supported by the KW even implicitly since the candidate answers are not extracted and thus not differentiated from other words frequently co-occurring with the words of the question.

Thus, from both algorithmic and user perspectives, *QA should be a better fit for a task involving fusion,* which is, of course, not surprising, because it was deliberately designed and built to accomplish the task. The above considerations also explain why the prior empirical studies could not find difference in performance between QA and KW: the tasks that were used by prior studies simply did not involve any fusion, thus QA was offering little additional benefit.



Chavez is a close ally of Cuba and Iran , whose president , Mahmoud Ahmadinejad , visited **Venezuela** on Saturday . Chavez said Venezuela and Iran agreed to push . . .
http://www.iol.co.za/index.php?set_id=1&click_id=122&art_id=qw1168757462763B215

. . . is the highest reached by Chavez , since the first months of his administration . . . . Recently , the Venezuelan leader visited **Argentina** and Brazil . . . .
http://www.change-links.org/chavez32005.htm

Chavez visited **Argentina** from Saturday to Wednesday in a semi official trip , in which signed cooperation programs with Argentinas President Nestor Kirchner . . .
http://english.pravda.ru/main/2001/07/23/10744.html

. . . revolutionary process undertaken by Chavez and the Venezuelan masses . . . . visited Saddam Hussein in **Iraq** (before the war) and Muâ ammar Gadhafi in Libya . . . .
http://www.nathanielturner.com/venezuelanrevolutionquestionsanswers.htm

Chavez emerged late Thursday from a meeting with Saddam , . . . Venezuelan President Hugo Chavez , on a groundbreaking visit to **Iraq** , has attacked US meddling . . .
http://english.people.com.cn/english/200008/11/eng20000811_47970.html

1) Hugo Chavez has visited Iraq , Iran and Libya . Because he is a friend of those nations , he is branded an enemy of the **United States** . . . .
http://www.blackcommentator.com/22/22_re_print_pr.html

In July , Chavez visited a number of countries that are discordant with the **United States** , including Iran and Belarus and signed a fighter aircraft import . . .
http://www.washingtonobserver.org/en/document.cfm?documentid=64&charid=3

Many here say they believe Chavez dreams of the day he can cut off the **United States** and sell to countries he considers more friendly . Chavez visited . . .
http://www.washingtonpost.com/ac2/wp-dyn/A35193-2005Mar14?language=printer

In 1923 24 (and again in 1926 28 , 1932 , and 1935 36) Chavez visited the **United States** . It excited him . The technical advance and invention manifest in the . . .
http://links.jstor.org/sici?sici=0027-4631(193610)22%3A4%3C435%3ACC%3E2.0.CO%3B2-Z

Here are the links to some incredible work by Chavez Parkside teachers and . . . 8th grade students visited the **United States** Supreme Court on March 23rd . . . .
http://www.cesarchavezhs.org/parkside?page=1

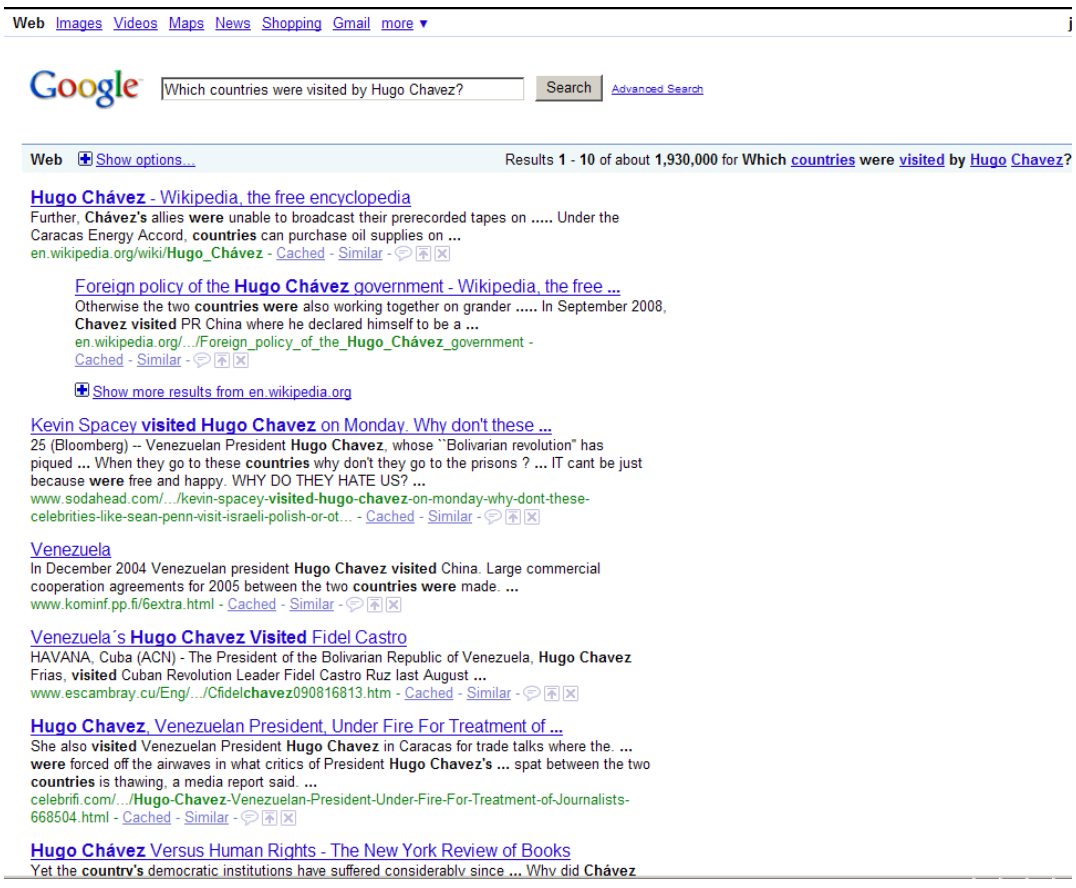
In 2000 Chavez visited other OPEC countries , becoming the first foreign head of state to visit **Iraq** since the 1991 Gulf War . He is close to President Fidel . . .
http://www.infoplease.com/ipa/A0108140.html

The visit by Chavez comes amid growing Venezuelan oil sales to **China** , . . . In China , Chavez is to meet with President Hu Jintao and Wen Jiabao , . . .
http://english.aljazeera.net/NR/exeres/E1EC64A1-8BA6-4B9C-AA65-151C9D9D2566.htm

In December 2004 Venezuelan president Hugo Chavez visited **China** . Large commercial cooperation agreements for 2005 between the two countries were made . . . .
http://www.kominf.pp.fi/6extra.html

Mi equipo 100%

Figure 1. WebQA Output for a List Question That Requires Fusion



Web Images Videos Maps News Shopping Gmail more ▾ j

Google Which countries were visited by Hugo Chavez? Search Advanced Search

Web Show options... Results **1** - **10** of about **1,930,000** for **Which countries were visited by Hugo Chavez?**

Hugo Chávez - Wikipedia, the free encyclopedia
Further, **Chávez's** allies **were** unable to broadcast their prerecorded tapes on **.....** Under the Caracas Energy Accord, **countries** can purchase oil supplies on **...**
en.wikipedia.org/wiki/**Hugo_Chávez** - Cached - Similar -

Foreign policy of the **Hugo Chávez** government - Wikipedia, the free ...
Otherwise the two **countries were** also working together on grander **.....** In September 2008, **Chavez visited** PR China where he declared himself to be a **...**
en.wikipedia.org/.../Foreign_policy_of_the_**Hugo_Chávez**_government - Cached - Similar -

Show more results from en.wikipedia.org

Kevin Spacey **visited Hugo Chavez** on Monday. Why don't these ...
25 (Bloomberg) -- Venezuelan President **Hugo Chavez**, whose ``Bolivarian revolution'' has piqued **...** When they go to these **countries** why don't they go to the prisons ? **...** IT cant be just because **were** free and happy. WHY DO THEY HATE US? **...**
www.sodahead.com/.../kevin-spacey-**visited-hugo-chavez**-on-monday-why-dont-these-celebrities-like-sean-penn-visit-israeli-polish-or-ot... - Cached - Similar -

Venezuela
In December 2004 Venezuelan president **Hugo Chavez visited** China. Large commercial cooperation agreements for 2005 between the two **countries were** made. **...**
www.kominf.pp.fi/6extra.html - Cached - Similar -

Venezuela´s **Hugo Chavez Visited** Fidel Castro
HAVANA, Cuba (ACN) - The President of the Bolivarian Republic of Venezuela, **Hugo Chavez** Frias, **visited** Cuban Revolution Leader Fidel Castro Ruz last August **...**
www.escambray.cu/Eng/.../Cfidel**chavez**090816813.htm - Cached - Similar -

Hugo Chavez, Venezuelan President, Under Fire For Treatment of ...
She also **visited** Venezuelan President **Hugo Chavez** in Caracas for trade talks where the. **...** **were** forced off the airwaves in what critics of President **Hugo Chavez's ...** spat between the two **countries** is thawing, a media report said. **...**
celebrifi.com/.../**Hugo-Chavez**-Venezuelan-President-Under-Fire-For-Treatment-of-Journalists-668504.html - Cached - Similar -

Hugo Chávez Versus Human Rights - The New York Review of Books
Yet the **country's** democratic institutions have suffered considerably since **...** Why did **Chávez**

Figure 2. Keyword (Google) Output for a List Question That Requires Fusion

III. THE RESEARCH PROTOTYPE

Knowledge "Light" Versus Knowledge "Heavy" Systems

Modern QA technologies rely on many components, including document retrieval, semantic analysis, syntactic parsing, and explanation generation. In order to answer such questions, a typical QA system would: (a) transform the user query into a form it can use to search for relevant documents (Web pages), (b) identify the relevant passages within the retrieved documents that may provide the answer to the question, and (c) identify the most promising candidate answers from the relevant passages. The QA systems are designed based on techniques from Natural Language Processing (NLP), Information Retrieval (IR), and Computational Linguistics (CL). For example, Falcon [Harabagiu et al., 2000], one of the most successful systems in TREC competitions, is based on a prebuilt hierarchy of dozens of semantic types of expected answers (person, place, profession, date, etc.), complete syntactic parsing of all potential answer sources, and automated theorem proving to validate the answers. Their work encompasses decades of elaborate linguistic modeling and manual resource building.

Those "knowledge heavy" NLP-based approaches have a strong advantage: they can be applied to smaller collections (e.g., corporate repositories, collections of e-mails) and still provide good performance. However, none of the known top performing systems has been made publicly open to the other researches for follow-up investigations, most likely because of the expensive knowledge engineering required to build such systems and the related intellectual property issues. As a result, it is still not known what components of these systems are crucial for their success, and how well their approaches would perform outside of the TREC test sets or outside the proprietors' laboratory.

On the other hand, the algorithms behind many of the nonlinguistic ("knowledge light") systems have been disclosed (e.g., Voorhees and Buckland, 2007) and are possible to replicate. In the standard tests, the performances of most of the redundancy/pattern-matching-based systems have been found comparable to each other [Voorhees and Buckland, 2007]. Their strengths/weaknesses with respect to specific question types was found to be similar. Knowledge light approaches typically reach 75 percent of the performance of the knowledge heavy approaches (as measured during TREC competitions), but their important advantage for empirical research is allowing replication.

By using the publicly available and replicable technologies within our prototype, *we have addressed one of the methodological challenges mentioned above in the Introduction section: the availability of the technology for testing*.

Prototype Description

Foundations

This section briefly overviews the "knowledge light" QA technology that is behind our prototype. The details can be found in prior work [Roussinov and Robles-Flores 2007; Roussinov et al. 2008]. We based the prototype on an existing QA system that is grounded in pattern matching and redundancy (e.g., Clarke et al., 2001). In terms of performance and algorithms, it is similar to others within the redundancy-based family of QA systems [Dumais et al., 2002; Ravichandran and Hovy, 2002; Roussinov and Robles-Flores, 2007], which rely on constructing and sending several queries to a commercial Web search engines (e.g., Google in our case) and analyzing the returned snippets. The major advantage of such "knowledge light" approaches is that the underlying algorithms are publicly disclosed, fairly simple to implement, and independent of any elaborate linguistic resources such as ontologies of questions, transformation rules, parsers, etc. The other important advantage of such open domain Web question-answering system is that it uses the same documents (Web pages) as potential answer sources as the search engine on top of which it operates, which is, in our case, Google. This makes empirical comparison of the two tools (QA and KW) more meaningful, since they draw from the same knowledge sources.

Although the implementation details vary, all the *redundancy*-based approaches take their roots in the automated learning (or manual construction) of the answer *patterns*. Although many variations of pattern language have been proposed, they are all essentially trying to capture the possible formulations of answers. For example, an answer to the question *"What is the capital of China?"* can be found in a sentence *"The capital of China is Beijing,"* which matches a pattern *\Q is \A,* where *\Q* is the target of the question (*"The capital of China"*) and *\A = "Beijing"* is the text that forms a *candidate answer. \A*, *\Q*, *\T*, *\p* (punctuation mark), *\s* (sentence beginning), \V (verb) and * (a wildcard that matches any words) are the only special symbols used in our pattern language. *\T* stands for optional semantic category of the expected answer, e.g., for the question *"In which city is the Eiffel Tower located?" \T = "city."*

Triangulation, a term which is widely used in intelligence and journalism, stands for confirming or disconfirming facts by using multiple sources. In order to employ the full power of triangulation, for each question (e.g., *Who is the CEO of IBM*?), each candidate answer has to be extracted from the sentences returned by answer services (e.g., *Samuel Palmisano* from the sentence *Samuel Palmisano became the twelfth CEO of IBM*), so that the answers can be compared with the other candidate answers (e.g., *Sam Palmisano*—a possible variation).

Finally, output sentences are re-ranked according to the expected number of correct answers contained, e.g. by a formula from Roussinov and Chau [2008]:

$$\text{score(S)} = \sum_{c(i) \in S} p(i) \qquad (1)$$

where *p(i)* is the probability of each candidate answer *c(i)* in the sentence *S* to be correct, which is approximated by the score of the candidate answer after the triangulation step mentioned above.

After receiving feedback from the preliminary user studies, we realized that this approach is not enough for the questions involving fusion where the algorithm needs to present several answers. The next section explains our novel solutions suggested for this challenge.

Promoting Diversity: Novel Re-ranking Based on Reported Evidence

This section presents our simple but novel algorithm promoting diversity in the reported answers. We introduce the notion of "sufficient reported evidence": once a certain candidate answer has already appeared enough many times in the output, it should be excluded from formula (1) above. To quantify this "sufficiency," we designed the following heuristic model. We evaluate the degree of evidence to which a given sentence supports a given candidate answer through linear mapping from the maximum possible lexical overlap between the question words and the words in the sentence into the [0,1] interval. For example, given the question "What countries have been visited by Hugo Chavez?" the sentence "Chavez likes Cuba" is estimated to provide only *12.5 percent* support for the candidate answer "Cuba," since there is only one word overlap "Chavez" out of maximum eight. The sentence "Hugo Chavez visited Mexico on his way to Cuba" provides 3/8 = *37.5 percent* support for both candidate answers "Mexico" and "Cuba."

Each time when a new *j*-th sentence is added to the output, the probabilities (scores) associated with the candidate answers that are used in formula (1) are discounted by the following adjustment: *p(i,j+1) = (1 – e) p(i,j)*, where *p(i,j)* is the (adjusted) probability of being correct after *j*-th sentence has been already reported. Before generating the snippets, *p(i,0)* is set to *p(i)* from the formula *(1).* After each snippet is reported, the remaining sentences are re-ranked according to the algorithm described in the preceding section and modified accordingly:

$$\text{score(S)} = \sum_{c(i) \in S} p(i, j) \qquad (2)$$

In the example above, this results in discounting of the candidate answer "Venezuela" after the first sentence has been reported, so the sentences containing "Argentina" receive the highest rankings. Once they are reported, "Iraq" becomes the most probable, yielding later to "United States," which in turn yields to "China." While some of those answers are wrong, the majority of them are correct. This illustrates that the algorithm successfully combines the diversity and the likelihood of being correct and, as a result, reports more correct answers than the KW output shown in Figure 2 for comparison.

IV. PROTOTYPE EVALUATION: OVERCOMING METHODOLOGICAL CHALLENGES

TTF as a Theoretical Framework

The review of the TTF literature suggested that its performance metrics are typically related to the concepts of *effectiveness, efficiency, and user satisfaction*. In general, *effectiveness* means that the user of the technology is able to successfully complete the task. *Efficiency* is a concept related to the way of being economical in terms of the use of resources to complete the task. *User satisfaction* is related to the perception of the user about the technology and how helpful it is perceived to be. The TTF literature presents different approaches, and in some cases it uses "utilization" as a proxy for performance.

While studying QA involving fusion (e.g., list questions), we suggest measuring the user performance by three indicators: (a) *mental workload*, (b) *user satisfaction*, and (c) the *recall*. The last measure originates from the information retrieval literature [Salton and McGill, 1983] and refers to the ratio of the relevant responses to the total number of all possible relevant responses. While precision is also often measured in the information seeking studies, we believe that it is not necessary in the empirical investigation of QA when the participants are expected to verify their answers, thus they serve as the judges of the correctness at the same time. The recall (percent of all the correct items that are found) becomes the ultimate metric (outcome measure) of performance, and is simply referred to as *performance* below.

It is important to distinguish the user performance measures listed above from the system measures, e.g., a number of correct answers included in the output. Even if the system performs well, the user still needs to be able to take full advantage of that, which at least necessitates (a) being able to recognize the answers and (b) to be sufficiently convinced that they are indeed correct. This makes the outcome of an empirical study comparing QA vs. KW at least not trivial and may explain why no empirical superiority of QA has been demonstrated so far.

The other metrics that we validated in the user study were the following:

1. *Mental effort*. This variable measures perceived mental workload on the user performing the task. This is a self-reported measure through the exit questionnaire (the NASA-TLX instrument). We measure mental effort separately for each tool being used.
2. *User satisfaction*. This is the perception of the user about the use of the technology and how it helps in finding information; therefore, an exit questionnaire was presented to the participants to measure satisfaction with each tool separately. The user satisfaction questions were adapted from Doll and Torkzadeh (1998).

For a study involving information seeking tasks, TTF would also suggest proceeding through the following two phases, not necessarily involving the same participants: (1) Task Creation and (2) Task Execution. The next two sections provide the details.

Task Creation

As explained in the previous chapter, to test the involved technologies, we needed tasks involving fusion. We limited the situation to a fairly common scenario in which the user needs to find the answers to a list question, for which aggregating a number of items from different sources (fusion) is needed. Because studying QA empirically is a nascent area of research, there are no well-developed standard tasks available, so they had to be created for the experiment and, preferably, before the system is tuned, to avoid over-fitting the system for those specific tasks. We decided not to use the questions from past TREC competitions because (a) we discovered that only a few of them required fusion, and (b) they were already used extensively to test our (and similar) prototype(s) during their implementation and in prior studies.

First, we identified and formalized the necessary characteristics for the tasks to be classified as fusion based on the literature discussed above. This resulted in the instructions that we gave to the "task creators," which can be briefly summarized as the following: "Think of 10 simple (no longer than 20 words, and not involving logical structures such as negation, conjunction or disjunction) questions, each asking for a list of items. The items should be named entities (people, organizations, dates, numbers, etc.). You are not expected to find all the answers yourself but should expect the total number to be approximately between 10 and 30. The answers to the questions should not be easily found on a single page, such as those in Wikipedia or other online resources." We provided positive and negative examples of questions with explanations.

As "creators of questions," we invited graduate students (master's or Ph.D. level) or recent (<2 years) graduates. People with a variety of areas of expertise were intentionally included in order to solicit questions from different knowledge domains: urban planning, social justice, information systems, management, biology, engineering, computer science, and psychology. The request to contribute was sent to forty-two individuals and nineteen responded with questions (45.2 percent response rate). A total of 179 questions were collected.

A brief inspection of the collected questions showed that the degrees of meeting the requirements set forth varied greatly and necessitated further cleaning of the set, which was performed by one of the coauthors of this article. We did not notice any significant ambiguity when deciding what questions fit the criteria that we gave to the contributors. The reasons for excluding questions were (in the order of being more common): (1) answers existing on a single Web page, e.g., all the answers for *How many species of Tetraedron exist?* in a single Wikipedia article, (2) too many answers existed, e.g., for *How many types of wood were used in the construction of Colon's ships?* (3) expected answers not being named entities, e.g., *What types of reptiles are confiscated annually at major airports*.

It's worth emphasizing that no attempts to run the questions through a QA system were made at this stage. After the verification process, thirty-six questions remained. Thus, as a byproduct of our user study, we created a test set of questions, which can be reused in future experiments. For the user study, we randomly selected twenty-eight questions (4 x 7), as a convenient number to perform pseudo-random balanced allocation of the participants in the experiment. Those questions are listed in Table 2.

Table 2: Questions Used in the User Study	
Which countries has Governor Janet Napolitano visited?	Which cities have St. John Boutique outlets?
What companies control low earth orbit satellites?	Which countries withdrew from OPEC?
What are the locations of the manned lunar landings?	What techniques exist to measure Fe isotopes?
What fresh food products are exported from the Chilean Los Lagos region?	What countries were members of the original League of Nations?
What types of chickens are raised in the United Kingdom?	Which actors are also authors?
Who are the contemporary composers of the medieval composer Moniot d'Arras?	Who are the first ladies of South America countries?
Which towns in Germany were bombed in World War II?	Which Nobel Prize winners were born in Latin America?
Which baseball stadiums offer tacos?	What types of reptiles are confiscated annually at major airports?
What Latin America presidents were prosecuted for corruption?	Which actors were taller than Leonardo DiCarpio in Titanic?
What unique minerals were found in underground crystal mines in the past 20 years?	Which countries launched military satellites?
Which Florida cities have Opera Houses?	With which African institutions has Arizona State University established agreements in the area of political sciences?
Which public companies in the U.S. have existed for more than 150 years?	What countries have Precambrian outcrops?
Which soccer players scored more than 500 goals?	Which biographers wrote biographies of Che Guevara?
What Klingon vessels did Captain Kirk order the Enterprise to attack?	Which capital cities were founded before Christ?

The important lessons learned from this phase were the following: (1) it is possible to create a test set of questions with necessary characteristics of fusion. (2) It is reasonable to expect some amount of cleaning of the suggested questions; thus other participants not related to the study may need to be recruited for this purpose, with the instructions created in advance.

Once the questions were selected, they have been processed by our Web QA prototype, so the answers would be stored in its cache. Otherwise, producing the answers during the user study would create unnecessary wait time. We also verified that the system was not only able to produce correct answers for most of the questions, but also produced two to three times as many correct answers on the first page compared with when the same question was entered verbatim to Google. This was not surprising since Google is not designed to answer list questions, and the capability to do so was found to be low in previous studies [Roussinov et al., 2008]. In comparison, *we were not able to observe the same difference in the returned correct answers for the sample of questions not involving fusion*, which we identified within the same original set of 179 questions, which supported our conjecture that the presence of fusion to warrant the use of QA is vital.

To further investigate if the users could actually take advantage of the good answers given to them by the system, or if they could alternatively do as well instead of creating effective Google queries themselves, we proceeded to a user study described in the next section.

V. TASK EXECUTION: A USER STUDY

We involved a convenience sample of 120 undergraduate students taking an introductory Information Systems course in a business school, motivated by a small extra credit.

The participants in the experiment had the following characteristics (information collected through an exit questionnaire): eighty-four male and fifty-five female (three did not respond). On average they have 2.13 years in college (fifty-four had one year of college, forty-four had two years, twenty-seven had three years, and fourteen reported more than three years of college). Regarding age, 134 participants reported their age between eighteen and thirty, while one reported below eighteen and six reported between thirty-one and fifty (one participant did not respond to this question). In the native language question, 123 reported being English-language native speakers and sixteen had other language as primary (one left the question blank). Regarding experience with search tools, 113 participants reported a daily use of search tools, while twenty-four reported using them several times per week, and four participants said they use them around once per week or less (one participant did not respond). On an "expertise" scale from 1 to 5 (1 = expert and 5 = inexperienced), twenty-six participants considered themselves in level 1, eighty in level 2, and thirty-three in level 3, while two considered themselves in levels 4 or 5 (one participant

did not respond to the question). Regarding use of e-mail, 130 participants responded that they use e-mail every day, nine use it several times per week, and two use it around once per week (one participant did not respond) The questions about the participants' use of the World Wide Web shows that 131 of them use it daily, nine several times per week, and two around once per week (one did not respond).

Each participant in this phase was assigned six tasks to perform (find answers to questions), alternating the use of both technologies (QA or KW), and the seventh task where he/she could choose which technology to use. The participants completed exit questionnaires. Considering that each participant was given five minutes per task and performed seven tasks, with the overhead for the instructions and filling out the questionnaires, the study took approximately one hour, which in the past was suggested as a reasonable time frame for which it is possible to recruit volunteer subjects.

Our instructions for using QA allowed switching to KW once all the system answers were checked. We decided not to block users' access to KW search when they were instructed to use QA because (1) that would mimic a real-life scenario better since these days users have access to keyword search engine portals virtually anytime and (2) using only QA would unfairly constraint searching for the answers, since the users did not have any control over the output of the QA system, so they would not be able to find any answers if the output was not useful for some reason (e.g., the system misinterpreting the question, not processing it correctly, etc.). Thus, throughout the rest of this article, the QA tool actually stands for the combination of question answering and traditional keyword searching (KW).

For this reason, it would seem intuitive to expect that the performance with QA should be at least as good as with KW, since the user approaching a task with QA always had an option to switch to KW. However, it is important to keep in mind, that the switching itself may take additional time and require cognitive effort. Indeed, the user still needs to sift through the output of the QA while trying to find any useful answers, and only after checking all of them or after giving up for some other reason, the user switches to KW.

Even having a large number of users involved in our study, we expected our methodology to be further refined. Thus, we treated our investigation as simply a "user" or "pilot" study to learn important lessons rather than to conduct a formal experiment.

Table 3 summarizes our hypotheses and the tests run. When the analysis was performed by subjects, the results happened to be opposite of what we expected: more users preferred KW, felt less cognitive load, and were more effective with it, than with QA. However, since the subjects used the same set of twenty-eight questions, it made sense to perform the analysis by questions as well, which showed that the hypothesized differences were not statistically significant; thus neither system can be judged by the experiment as uniformly better than the other: there was strong dependence on a particular question; thus a larger sample of question would be needed to support the claim that one system is better than the other.

Table 3: Summary of Hypotheses

Hypotheses	Analysis
H1: Utilization	Proportions test
H2a: Mental Effort	Paired t-test
H2b: User Satisfaction.	Paired t-test
H2c: Performance	Repeated measures ANOVA

Figure 3 below shows the mean scores obtained for each question, regardless of the technology used. The conclusion is that there may be an effect by the question number. This implies that some questions are more or less difficult than others.

As Table 4 illustrates, for the same question, the average across-subjects score was statistically significantly higher when the subjects used the technology of their choice (while working on task 7) rather than when they were asked to use a specific one. This leads to several important conclusions: (1) After gaining some experience with using both technologies and with the tasks in the experiment, the participants were able to correctly judge which technology is better suited to answer a specific question. (2) Still a substantial proportion of participants (35 percent) chose to use QA, which shows it as a promising new technology. (3) By making their choice, the participants were able to improve their performance. We believe all of these findings are encouraging from the point of view of investing in the new emerging QA technology.

Communications of the Association for Information Systems

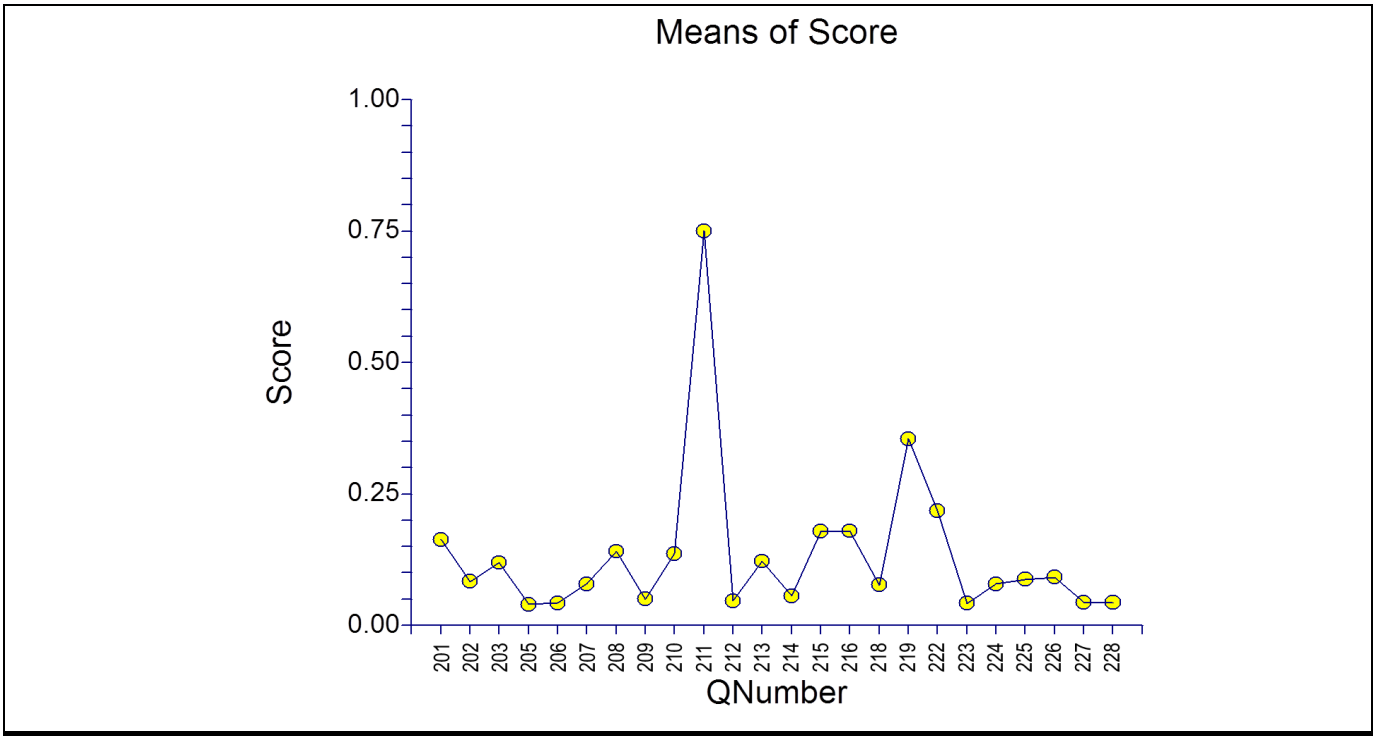


Figure 3. Average Score for Each Question Regardless of Technology Used

Table 4: Mean Performance Scores of Participants Based on their Choice of Technology. The Differences Are Statistically Significant (alpha = 0.05).

	Web keyword searching	Web question answering
Performance when assigned	0.11812385	0.09187134
Performance when chosen	0.13405913	0.11031367

While the participants in our preliminary studies and the questions creators were mostly graduate students, the participants in this phase were undergraduates. This may explain why a much *higher than expected proportion of erroneous answers were reported* (50 percent vs. 5 percent). However, follow-up interviews with some of the users suggested that, when verifying the answers was adequately emphasized, the users were able to easily discard erroneous answers; thus our conjecture on the possibility of using recall as the primary performance metric was still supported. In addition to clarification and training in verifying the answers, future empirical designs may also involve penalties for reporting wrong answers. The additional explanation of this difference may be the time pressure that the users experienced. Since the experiment instructions expected not only to find the answers, but also to verify the correctness of them, more time was needed for each task. The fit between the technology and the task was decreased. Future empirical designs should allow sufficient time and motivation for the users to simulate the real-life tasks when the users have vested interest in the correctness of the reported answers. This would provide a better fit between the technology and the task.

The number of correct answers returned by the participants using the QA system correlated positively with the number of answers returned by the system. This highlights the (rather expected) importance of the system to do its best in interpreting the question and reporting as many correct answers as possible. Some questions, especially those requiring understanding the subtleties of human language, happened to be more difficult. The ANOVA results on the performance showed large interaction effect in which the question number has an effect on the scores obtained. The variability of user scores by questions suggests that for a full-scale experiment *a substantially larger number of questions is needed*.

The tasks (questions) happened to be difficult for the participants with using either system. As the post-questionnaires indicated, the participants were overall unsatisfied with their performance and often felt frustrated with their overall ability to accomplish the assigned tasks. The proportion of attempts (task/participant pairs) in which any answers were reported was 52 percent, while the proportion of attempts resulting in any correct answers was only 31 percent. Using QA was new for the participants and demanded even more effort simply to get familiar with it and follow the experiment instructions. The more familiar tool (Google) caused less frustration and imposed smaller mental load.

It is interesting to note that the QA system output contained several more correct answers than what the participants actually found. It seems as if the participants did not "see" these answers in the output. Table 5 provides more details on this. It is possible that, due to time constraints in the experiment, the participants were not able to spend enough time to review the output carefully and to follow the links when needed.

No major issues were observed regarding the lack of understanding of or not following the instructions. We found the technology involved was stable enough and did not experience any technical glitches during the user study.

Table 5: Comparing the Number of Items Found for Each Question (Task) and the Number of Items in the WebQA Output

Question	Items found in the WebQA output	Average number of items found by participants assigned to QA
201	3	1.26666667
202	8	1
203	0	1.5
205	3	0.13333333
206	6	0.66666667
207	2	1.76470588
208	5	1
209	1	1.4117647
210	7	2.52941176
211	2	1.85714286
212	7	0.46153846
213	5	3
214	7	2.8
215	1	1.4
216	6	2.67
218	0	0.2
219	12	5.5
222	1	1.5
223	11	1.6
224	8	2.8
225	3	2.3
226	3	0.1
227	0	0
228	3	1.8

V. CONCLUSIONS

By using Task Technology Fit (TTF) theory, we have been able to explain why online question-answering (QA) technology has not been so far demonstrated to be superior to traditional keyword search (KW). We have suggested that, when the task does not require fusion, QA offers little benefit in addition to KW. However, when the fusion is expected, as in the case of finding answers to list questions, QA could have an impact on the results. It may be possible that the most popular questions do not require fusion, since somebody has already combined all the answers for them in a single source. Thus, QA may be a more suitable technology for less "main-stream" questions.

We have also demonstrated how Task Technology Fit theory can successfully guide designing the methodology to test this impact and have validated our methodology through a user study, which provided a number of valuable lessons for a future full-scale experiment.

We have described our novel prototype (a new system capable of answering list questions such as *What companies own low orbit satellites*?) specially designed and implemented for an evaluation experiment. Our prototype illustrates how publicly available "knowledge-light" building blocks can be used to overcome the limitations of the "knowledge-heavy" proprietary technologies and to allow replication outside the proprietors' laboratory.

A number of limitations were also discussed throughout our article, and we are going to address them in future research when preparing a larger scale laboratory experiment or a field study. The most important lesson learned is that more training in the use of QA tool is vital. Specifically, the user needs to learn how to quickly glance through the answers presented by the system, how to recognize and verify the most promising ones, how to verify the correctness of the answers found, and when to consider the task completed.

When the users had a choice of technology (on their last question), they scored better on average than the other users who answered the same question when being constrained to use a specific tool. This indicates that the users accumulated enough experience during the study to properly judge which tool is more appropriate for them to use on the given question and to use it effectively.

Our research makes an impact in the following ways:

1. It contributes to the theory by exploring the applicability of TTF to open domain question answering.
2. It presents and tests the methodology to evaluate information seeking tasks.
3. Our study provides insights to managers on whether it is worth investing in QA technologies and training the personnel, which, as we argue, greatly depends on the amount of fusion within the information-seeking tasks that the company encounters.

We certainly believe that this study will contribute to the managerial awareness of the capabilities and limitations of the modern question-answering technology within the overall information supply chain and the appreciation of the challenges of its methodological evaluation.

REFERENCES

Editor's Note: The following reference list contains hyperlinks to World Wide Web pages. Readers who have the ability to access the Web directly from their word processor or are reading the article on the Web, can gain direct access to these linked references. Readers are warned, however, that:

1. These links existed as of the date of publication but are not guaranteed to be working thereafter.
2. The contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.
3. The author(s) of the Web pages, not AIS, is (are) responsible for the accuracy of their content.
4. The author(s) of this article, not AIS, is (are) responsible for the accuracy of the URL and version information.

Alexa Top Sites (2010), http://www.alexa.com/topsites (current Feb. 24, 2010).

Brin, S., and L. Page (1998) "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems* (30)1–7, pp. 107–117.

Case, D.O. (2002) *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior, 1st edition,* Amsterdam, Netherlands: Academic Press.

Chen, C.C., J. Wu, and S.C. Yang (2006) "The Efficacy of Online Cooperative Learning Systems: The Perspective of Task-technology Fit," *Campus-Wide Information Systems* (23)3, pp 112–127.

Chung, W., H. Chen, and J. Nunamaker (2005) "A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration," *Journal of Management Information Systems* (21)4, pp. 57–84.

Clarke, C., G. Gormack, and T. Lyman (2001) "Exploiting Redundancy in Question Answering" in *Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, New Orleans, LA.

Clayburn, T. (2005) "Web Search for Tomorrow," *Information Week,* Mar. 28, 2005.

CNN Money (2011) "IBM's Jeopardy Supercomputer Beats Humans in Practice Bout," http://money.cnn.com/2011/01/13/technology/ibm_jeopardy_watson/index.htm (current Jan. 14, 2011).

Dishaw, M.T., and D.M. Strong (1998a) "Assessing Software Maintenance Tool Utilization Using Task-technology Fit and Fitness-for-use Models," *Journal of Software Maintenance-Research and Practice* (10)3, pp. 151–179.

Dishaw, M.T., and D.M. Strong (1998b) "Supporting Software Maintenance with Software Engineering Tools: A Computed Task-Technology Fit Analysis," *Journal of Systems and Software* (44)2, pp. 107–120.

Dishaw, M.T., and D.M. Strong (1999) "Extending the Technology Acceptance Model with Task-Technology Fit Constructs," *Information & Management* (36)1, pp. 9–21.

Doll, W.J., and G. Torkzadeh (1988) "The Measurement of End-User Computing Satisfaction," *MIS Quarterly* (12)2, pp. 259–274.

Dumais, S., et al. (2002) "Web Question Answering: Is More Always Better?" in *Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, Tampere, Finland.

Goodhue, D.L. (1998) "Development and Measurement Validity of a Task-Technology Fit Instrument for User Evaluations of Information Systems," *Decision Sciences* (29)1, pp. 105–138.

Goodhue, D.L. (1995) "Understanding User Evaluations of Information Systems," *Management Science* (41)12, pp. 1827–1844.

Goodhue, D., R. Littlefield, and D.W. Straub (1997) "The Measurement of the Impacts of the IIC on the End-Users: The Survey," *Journal of the American Society for Information Science* (48)5, pp. 454–465.

Goodhue, D.L., and R.L. Thompson (1995) "Task-Technology Fit and Individual-Performance," *MIS Quarterly* (19)2, pp. 213–236.

Goodhue, D.L., B.D. Klein, and S.T. March (2000) "User Evaluations of IS as Surrogates for Objective Performance," *Information & Management* (38)2, pp. 87–101.

Harabagiu, S., et al. (2000) "Falcon: Boosting Knowledge for Answer Engines" in *Text REtrieval Conference (TREC 9)*, Gaithersburg, MD.

Hevner, A.R., et al. (2004) "Design Science in Information Systems Research," *MIS Quarterly* (28)1, pp. 75–105.

Katz, B., et al. (2004) "Answering Multiple Questions on a Topic from Heterogeneous Resources" in *Text REtrieval Conference (TREC 2004)*, Gaithersburg, MD.

Klopping, I.M., and E. McKinney (2004) "Extending the Technology Acceptance Model and the Task-Technology Fit Model to Consumer E-Commerce," *Information Technology, Learning, and Performance Journal* (22)1, pp. 35–48.

Lyman, P., and H.R. Varian (2001) "The Democratization of Data," *Harvard Business Review* (79)1, pp. 137–139.

Marchionini, G. (1995) *Information Seeking in Electronic Environments, 1st edition,* Cambridge, U.K.: Cambridge University Press.

Maybury, M., et al. (2003) *New Directions in Question Answering: Papers from 2003 AAAI Spring Symposium, 1st edition*, Menlo Park, CA: American Association for Artificial Intelligence.

Mills, E. (2005) "Google to Yahoo: Ours Is Bigger," in *CNET News.com*, 2005.

Ravichandran, D., and E. Hovy (2002) "Learning Surface Text Patterns for a Question Answering System" in *Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA.

Robles-Flores, J.A. (2009) "Web Question Answering Technology: An Empirical Test of the Task-Technology Fit Model," Unpublished Ph.D. Dissertation. W.P. Carey School of Business, Arizona State University, Tempe, AZ.

Roussinov, D., and M. Chau (2008) "Combining Information Seeking Services into a Meta Supply Chain of Facts," *Journal of the Association for Information Systems* (9)3, pp. 175–199.

Roussinov, D., W. Fan, and J.A. Robles-Flores (2008) "Beyond Keywords: Automated Question Answering on the Web," *Communications of the ACM (CACM)* (51)9, pp. 60–65.

Roussinov, D., and J.A. Robles-Flores (2007) "Applying Question Answering Technology to Locating Malevolent Online Content", *Decision Support Systems* (43)4, pp. 1404–1418.

Salton, G., and M.J. McGill (1983) *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill.

Staples, D.J., and P. Seddon (2004) "Testing the Technology-to-Performance Chain Model," *Journal of Organizational and End User Computing* (16)4, pp. 17–36.

Voorhees, E., and L.P. Buckland (2007) *Proceedings of the Seventeenth Text Retrieval Conference TREC 2007.* Gaithersburg, MD.

Wilson, T.D. (2000) "Human Information Behavior," *Informing Science* (3)2, pp 49–55.

Zigurs, I., and B.K. Buckland (1998) "A Theory of Task/Technology Fit and Group Support Systems Effectiveness," *MIS Quarterly* (22)3, pp. 313–334.

Zigurs, I., et al. (1999) "A Test of Task-Technology Fit Theory for Group Support Systems," *The DATA BASE for Advances in Information Systems* (30)3–4, pp. 34–50.

ABOUT THE AUTHORS

José Antonio Robles-Flores is an Assistant Professor of Information Systems at the Graduate School of Business at ESAN University in Lima, Perú. He is particularly interested in testing new technologies related to information retrieval, knowledge acquisition, and knowledge transfer. He is also interested in studying how information and communication technologies (ICT) impact on organizations of developing countries. Dr. Robles-Flores has published in journals such as *Decision Support Systems* and *Communications of the ACM*; and in proceedings of conferences such as AMCIS, HICSS, WWW, JCDL, TREC and SIGIR. José Antonio earned his Ph.D. in Information Systems from W.P. Carey School of Business at Arizona State University. Before joining Academia, he worked as a software engineer and also managed Internet projects.

Dmitri Roussinov is a Senior Lecturer at the Department of Computer and Information Sciences, University of Strathclyde. He received his Ph.D. in MIS from the University of Arizona and has a prior MA degree in Economics from Indiana University and a diploma with honors in Computer Science from Moscow Institute of Physics and Technology, Russia. Prior to joining Strathclyde, Dr. Roussinov was an Assistant Professor at Arizona State University, and before that he served two years on the faculty at Syracuse University, School of Information Studies. His research interests include security informatics, applications of artificial intelligence to knowledge management, group decisions support systems, and electronic commerce. Dr. Roussinov has published in *IEEE Transactions on Knowledge and Data Engineering, Communications of the ACM, Decision Support Systems, and Information Processing and Management*. He has also presented at many well-known international conferences.

Copyright © 2012 by the Association for Information Systems. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than the Association for Information Systems must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or fee. Request permission to publish from: AIS Administrative Office, P.O. Box 2712 Atlanta, GA, 30301-2712, Attn: Reprints; or via e-mail from ais@aisnet.org.

Communications of the Association for Information Systems

ISSN: 1529-3181

EDITOR-IN-CHIEF
Ilze Zigurs
University of Nebraska at Omaha

AIS PUBLICATIONS COMMITTEE

Kalle Lyytinen Vice President Publications Case Western Reserve University	Ilze Zigurs Editor, *CAIS* University of Nebraska at Omaha	Shirley Gregor Editor, *JAIS* The Australian National University
Robert Zmud AIS Region 1 Representative University of Oklahoma	Phillip Ein-Dor AIS Region 2 Representative Tel-Aviv University	Bernard Tan AIS Region 3 Representative National University of Singapore

CAIS ADVISORY BOARD

Gordon Davis University of Minnesota	Ken Kraemer University of California at Irvine	M. Lynne Markus Bentley University	Richard Mason Southern Methodist University
Jay Nunamaker University of Arizona	Henk Sol University of Groningen	Ralph Sprague University of Hawaii	Hugh J. Watson University of Georgia

CAIS SENIOR EDITORS

Steve Alter University of San Francisco	Michel Avital Copenhagen Business School	Jane Fedorowicz Bentley University	Jerry Luftman Stevens Institute of Technology

CAIS EDITORIAL BOARD

Monica Adya Marquette University	Dinesh Batra Florida International University	Indranil Bose Indian Institute of Management Calcutta	Thomas Case Georgia Southern University
Evan Duggan University of the West Indies	Andrew Gemino Simon Fraser University	Matt Germonprez University of Wisconsin-Eau Claire	Mary Granger George Washington University
Åke Gronlund University of Umea	Douglas Havelka Miami University	K.D. Joshi Washington State University	Michel Kalika University of Paris Dauphine
Karlheinz Kautz Copenhagen Business School	Julie Kendall Rutgers University	Nelson King American University of Beirut	Hope Koch Baylor University
Nancy Lankton Marshall University	Claudia Loebbecke University of Cologne	Paul Benjamin Lowry City University of Hong Kong	Don McCubbrey University of Denver
Fred Niederman St. Louis University	Shan Ling Pan National University of Singapore	Katia Passerini New Jersey Institute of Technology	Jan Recker Queensland University of Technology
Jackie Rees Purdue University	Raj Sharman State University of New York at Buffalo	Mikko Siponen University of Oulu	Thompson Teo National University of Singapore
Chelley Vician University of St. Thomas	Padmal Vitharana Syracuse University	Rolf Wigand University of Arkansas, Little Rock	Fons Wijnhoven University of Twente
Vance Wilson Worcester Polytechnic Institute	Yajiong Xue East Carolina University		

DEPARTMENTS

Information Systems and Healthcare Editor: Vance Wilson	Information Technology and Systems Editors: Dinesh Batra and Andrew Gemino	Papers in French Editor: Michel Kalika

ADMINISTRATIVE PERSONNEL

James P. Tinsley AIS Executive Director	Vipin Arora CAIS Managing Editor University of Nebraska at Omaha	Sheri Hronek CAIS Publications Editor Hronek Associates, Inc.	Copyediting by S4Carlisle Publishing Services

Communications of the Association for Information Systems