2005

How Question Answering Technology Helps to Locate Malevolent Online Content

Dimitri Roussinov
Arizona State University, dimitri.roussinov@asu.edu

Jose A. Robles-Flores
Arizona State Universaity/ESAN University, jose.robles@asu.edu

Follow this and additional works at: http://aisel.aisnet.org/amcis2005

How Question Answering Technology Helps
to Locate Malevolent Online Content

Dmitri Roussinov
Arizona State University
Dmitri.Roussinov@asu.edu

Jose A. Robles-Flores
Arizona State University / ESAN University
Jose.Robles@asu.edu

ABSTRACT

The inherent lack of control over the Internet content resulted in proliferation of online material that can be potentially detrimental. For example, the infamous "Anarchist Cookbook" teaching how to make weapons, home made bombs, and poisons, keeps re-appearing in various places. Some websites teach how to break into computer networks to steal passwords and credit card information. Law enforcement, security experts, and public watchdogs started to locate, monitor, and act when such malevolent content surfaces on the Internet. Since the resources of law enforcement are limited, it may take some time before potentially malevolent content is located, enough for it to disseminate and cause harm. Currently applied approach for searching the content of the Internet, available for law enforcement and public watchdogs is by using a search engine, such as Google, AOL, MSN, etc. We have suggested and empirically evaluated an alternative technology (called automated *question answering* or *QA*) capable of locating potentially malevolent online content. We have implemented a proof-of-concept prototype that is capable of finding web pages that may potentially contain the answers to specified questions (e.g. "How to steal a password?"). Using students as subjects in a controlled experiment, we have empirically established that our QA prototype finds web pages that are more likely to provide answers to given questions than simple keyword search using Google. This suggests that QA technology can be a good replacement or an addition to the traditional keyword searching for the task of locating malevolent online content and, possibly, for a more general task of interactive online information exploration.

Keywords
Information systems security, information retrieval, question answering, world wide web.

INTRODUCTION

After the September 11 attacks, the world started to pay close attention to its vulnerable assets, one of which is undoubtedly the Internet -- the backbone of modern information superhighway. Making cyber infrastructure resilient to any attacks or misuse became a priority for scientists and funding agencies (National Science Foundation, 2003). However, the media still reports numerous cases of government web sites "defaced" or shut down by hackers (Swartz, 2004). In addition, the proliferation of illegal spamming, computer viruses, identity theft, software piracy and fraudulent schemes has threatened the trust behind electronic means of communication to the degree of becoming a threat to the national cyber infrastructure (Verton and Verton, 2003).

Due to the inherent lack of control over the Internet content, the techniques that cyber criminals use, are easily available, and very often do not require more than average expertise or skills. A recent example reported in popular press (Jacques, 2004) is the shadowcrew.com web site, which acted as a one-stop shop for false documents and information stolen by hacking into computers. The site also posted hacking methods and maintained a forum for hackers to exchange their ideas and techniques.

We define the content that can facilitate committing crimes (not necessary in cyberspace) as *malevolent*. The shadowcrew.com example clearly illustrates that *malevolent content frequently co-exists with illicit activities*. Law enforcement and public watchdogs actively locate, monitor, and sometimes act on such content (Weimann, 2004). For example, the "sting" operation performed by US Secret Service on shadowcrew.com resulted in 28 arrests across different countries (Jacques, 2004). Many similar web sites have also been forced to close down under political pressure (Weimann, 2004).

However, while the most notorious sites are acted upon, many less known ones can remain undetected (Reid et al., 2004). Since the resources of law enforcement and researchers are limited, it may take some time before potentially malevolent content is located, enough for it to disseminate and cause harm. One practical way to search the content of the Internet, available for law enforcement and public watchdogs, is by using a search engine, such as Google, AOL, MSN, etc. The

agents need to come up and maintain a list of keywords that would potentially uncover the parts of "Dark Web", run several queries and laboriously analyze the search results (Reid et al., 2004).

The algorithms that search engines use are based on lexical match (so called "bag of words" approach) in which the pages are represented as sets (bags) of words. This approach is known to result in a very well known problem of information overload on the Web (Lyman and Varian, 2000; Roussinov and Chen, 2001; Turetken and Sharda, 2004) considering that the web has more than 4 billion pages, the vast majority of them legitimate and harmless. Performing a Google (or other search engine) search on the topic of "hacking" and "phishing" results in the thousands of pages from the news media (e.g. cnn.com) or political discussion forums (e.g. soc.culture.usa) since the search engines' algorithm locate the content based on the lexical match and the popularity of the web sites (Brin and Page, 1998), thus mostly overlooking "shady" portions of the web. It is the level of technical detail that can distinguish innocuous pages from harmful ones (e.g. news from "how-to" manuals) but this level of depth cannot be picked by lexical matching or link analysis.

A recently emerging alternative to keyword based web searching is automated question answering (QA). A typical QA system would take the question in a natural language such as "How can I guess a password?" and return an answer such as "You can use a password dictionary to guess passwords." and/or a link to a source page that provides the answer. Recent breakthroughs in the QA technologies have been reported (Voorhees and Buckland, 2003) and are briefly reviewed in the next section.

Several research groups have made publicly available their demonstration systems capable of finding the answers on the World Wide Web: Language Computer (www.languagecomputer.com), AnswerBus (www.answerbus.com), NSIR (http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi) and ASU QA Demo (http://qa.wpcarey.asu.edu) and attracted attention from the media. For example, Information Week recently mentioned ASU QA system as one of the most promising directions in the "Search of Tomorrow." (Claburn, 2005)

We believe that QA technology may be a better candidate for locating malevolent online content or, at least, be a good addition to the keyword searching, since it seeks the pages that may provide answers to the questions entered by law enforcement agents or public watchdogs, e.g. "How to break into a network?", "How to steal a credit card number?", etc. This accurate pinpointing should be more effective than lexical match and popularity analysis currently used by commercial search engines.

In spite of recent breakthroughs in QA technology and the promise that it carries for a number of applications, no experiments have been performed to compare Web QA tools and popular search engines in which human users apply the tools to accomplish search tasks. This is true for the more general domain of information exploration, and to the more specialized domain of security applications. As a result, *it is not entirely clear if and by how much QA technology can help locating malevolent online content in addition to the currently available keyword searching.* This lack of empirical comparison and the desire to provide law enforcement with more effective tools to locate and monitor malevolent content have motivated our study. Thus, our contribution reported in this paper is comparing two classes of technologies capable of locating potentially malevolent online content: 1) popular keyword searching, currently widely used by law enforcement and general public, and 2) emerging question answering (QA) technology. Our QA prototype, specially created for this study, exemplifies the second approach, while Google search engine represented the first. Up to our knowledge, no prior work has attempted this kind of comparison.

The next section overviews the technology behind our prototype. The empirical evaluation section follows. Then, we present our conclusions.

TECHNOLOGY OVERVIEW

In our study, we used the system presented by Roussinov and Robles-Flores (2004), available on the Web as ASU Question Answering Demo (http://qa.wpcarey.asu.edu/). Roussinov and Robles-Flores expanded work by Dumais et al. (2002) by automated identification and training of patterns, triangulation, and using trainable semantic filters instead of manually created ones. When tested on standard factoid questions, their system demonstrated comparable performance to that of Dumais et al. (2002) and other publicly available demonstrational prototypes.

Roussinov and Robbles-Flores (2004) approach begins with a natural language question, such as "Who is the CEO of IBM?" The software then matches the question against question patterns that it knows to create an initial query. For example, the pattern "who is" indicates that the answer will be a person. Using that formulation, the software looks at stored answer patterns. In this case, the software looks for the word pattern "The CEO of IBM" when it occurs near "is", to create a final query that is sent out to search engines. The candidate answers (matching the specified patterns) are then triangulated (confirmed or disconfirmed) against each other. Triangulation, a term widely used in intelligence and journalism fields,

stands for confirming or disconfirming facts using multiple sources. Their novel triangulation algorithm can be demonstrated by the following intuitive example. Imagine that we have two candidate answers for the question *"What was the purpose of the Manhattan Project?"*: 1) *"To develop a nuclear bomb"* 2) *"To create a nuclear weapon."* These two answers should reinforce (triangulate) each other since they are semantically similar; however, a straightforward frequency count approach would not pick this similarity. The advantage of triangulation over simple frequency counting is higher for less "factual" questions, such as those that may allow variation in the correct answers.

Pattern based approach has additional advantages over deep NLP approaches for locating content on the Web since it can look for grammatically irregular sentences or combinations of headings followed by answer paragraphs. For example, when the system is given the question "How to hack into computer networks" it looks for such patterns as "Re: How to hack into computer networks", "How to hack into computer networks Tutorial," "Introduction into hacking into computer networks," etc. Figure 1, reproduced here from Roussinov and Robles-Flores (2004), illustrates steps involved in training their system and answering questions. We used their system without any noteworthy modifications besides manually adding patterns to answer "How to" types of questions.

Using the system by Roussinov and Robles-Flores (2004) also offered the following advantages:

1) Their system is entirely transparent in the sense that all algorithms are detailed in the prior publication. It does not depend on manual tuning, is completely trainable from examples, thus the results can be easily reproduced.

2) It allows searching the entire Web, which is essential in our study, as opposed to using a system like AskJeeves that delivers answers only from a specially constructed database.
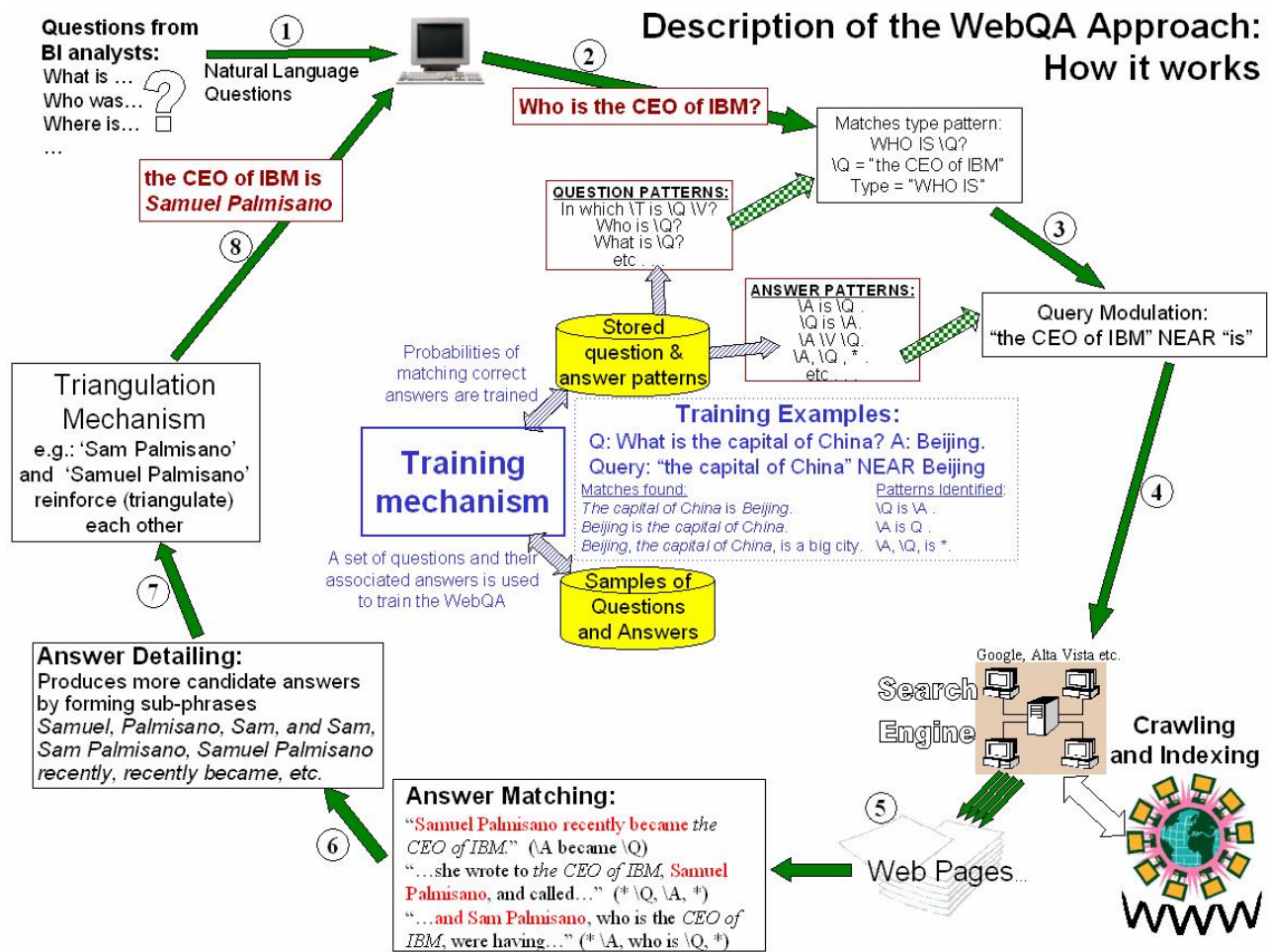

Figure 1. The WebQA approach (adapted from Roussinov and Robles-Flores, 2004)

EMPIRICAL COMPARISON

Research Questions and Methods

Our main objective was to compare two different approaches to locating web content (keyword search vs. pattern based question answering) exemplified by the following two tools: 1) our QA tool and 2) Google search portal. Since our QA tool also relies on Google for locating promising web pages that may contain an answer to a given question, in effect, we compared if our pattern matching QA layer was improving Google's ability to locate the specified content. Thus, our central ("task level") research question was the following:

Q1: *Does question-answering (QA) technology enhance the ability to locate potentially harmful content when compared to traditional keyword search?*

Our second research question was comparing the tools in an indirect way:

Q2: *Does using question-answering (QA) technology improve the relevance of the retrieved pages?*

The relevance of the page is determined by whether it helps to answer a given question or not, as operatioanalized below. The important difference between our two research questions is that even when the answer to Q1 is positive, the answer to Q1 still may be indecisive due to such other factors as (1) responsiveness of a system and (2) the users' ability to locate the desired content within the retrieved pages. To minimize the impact of the factor (2) in our study, we deliberately implemented our QA system in such a way that it presents results in the same way as Google does: it shows snippets of the pages instead of answer sentences (shown in Figure 2).

To address the above two research questions, we have performed our study in two phases accordingly. During the first ("task level") phase, the study volunteers came up with their own questions and attempted to find the answers using one of the tools, once for each question. During the second (blind evaluation) phase, the same volunteers only evaluated the results retrieved by each of the tools without knowing which particular tool had retrieved them. While the first phase was closer to a realistic scenario, it still included more random factors such as subjects' perception of the stopping condition (when answer was found) and responsiveness of each system (time). That is why we were not relying on analysis of the measurements collected exclusively from the phase one. Instead, we primarily viewed our first phase as a process of collecting a set of test questions needed for the second phase. The second phase excluded the factors of responsiveness (time) and allowed to repeat same questions with different tools, while significantly decreasing the variability in the measurements and making the design more efficient.

Metrics

Reciprocal Answer Rank

In order to evaluate the relevance of the retrieved pages we analyzed the search logs and computed the very popular metric of *reciprocal rank* of the first web page that can be considered an answer (Voorhees and Harman, 2003). Users decided themselves if the found page could be considered an "answer page" based on our criteria explained in the instructions: as long as the page helps to answer the question even "a little bit" it can be considered an answer. The users did not have to find an exact answer nor to compile the answers based on multiple sources. The users were also not expected to find all the answers (relevant pages), since that would prolong the experiment beyond the acceptable period of time. Since a reciprocal rank of the first answer typically correlates with other metrics such as average precision or recall, and the results were averaged across multiple subjects and questions, we believe this simplification was both valid and desirable.

Each time, when the user re-formulated Google queries, we increased the rank by 10 (the number of pages returned by Google as response to a query). The assumption behind this was that the user has at least glanced at all the top 10 snippets (or pages), did not find the answer, and moved along to a different query. We did not observe subjects reformulating their questions while using a QA tool that is why we did not need to decide whether to apply the same penalty.
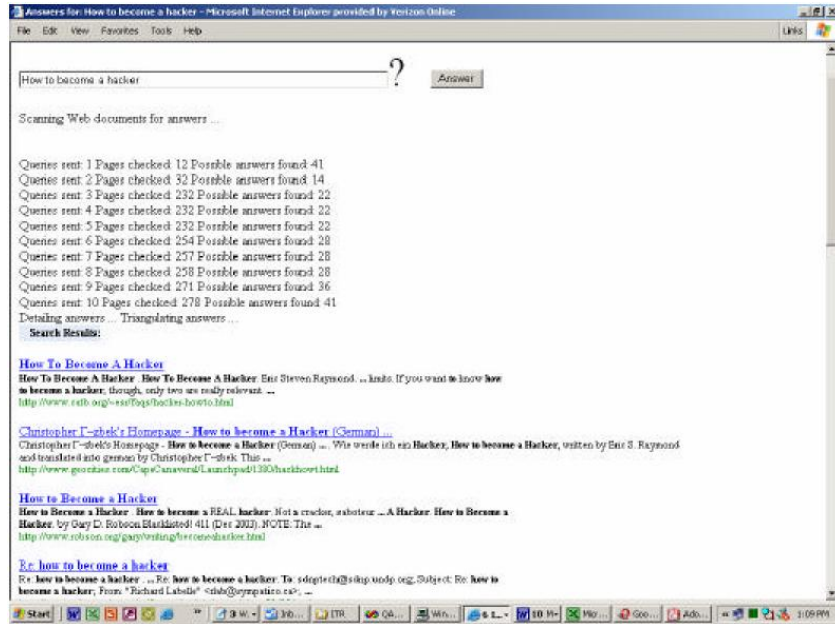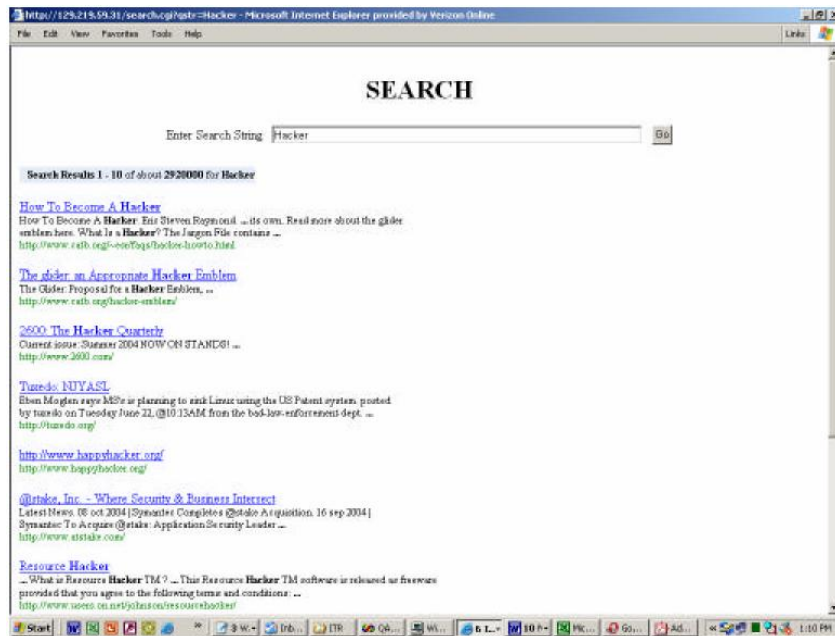
Figure 2. Question Answering session

Figure 3. Front end to Google

Testing the Instructions by a Pilot Study

Figure 2 shows a typical QA session. The interface is very straightforward: the user enters the question and receives a set of answers along with the links to the pages where the answers were found.

Since we only needed Google's functionality to provide keyword search, we had implemented a front end to Google that limits its capabilities to keyword search only, disabling all the other potentially distracting features at the portal such as image

search, toolbar icons, shopping, news, advertisement, etc. Our front end to Google (shown on Figure 3) forwarded the query to the search engine intact and presented the returned snippets in the same order and without any modifications either. All queries and their results were automatically logged.

For our pilot study, we asked 3 student volunteers to put themselves in the shoes of a "cyber-criminal" e.g. as if they were trying to commit illicit actions (e.g. "hack" into computer networks) and come up with 6 questions, answering to which would help to learn the criminal techniques. In order to get familiar with the topic and get inspired, we suggested spending 10 minutes searching Google News for topics related to cyber crime and skimming through news articles. Then, we asked the users to find answers to their questions using Google for 3 questions and QA tool for the other 3 questions, switching turns to minimize learning effects.

Task Level Phase
We involved 9 volunteers in our (task level) phase. The volunteers followed the instructions (slightly revised after our pilot study), online, on their own (not monitored), at the time of their choice. The volunteers were (all but one) undergraduate students in a school of business in a US research-type university, majoring in Information Systems, familiar with web searching and the domain of study (cyber crime).

Table 1 (posted at http://www.public.asu.edu/~droussi/IEEE2005-short-tables.html) shows the results of this phase. "NF" indicates that the answer was not found by the user within the time allowed (5 minutes), and was assigned the reciprocal rank of 0. We wanted to keep a user session within a limit of one hour, thus each user tried only 3 different questions with each tool (6 questions total). This created high variability of assessment within our relatively small sample, which explains why we did not find any statistically significant difference in user preference with respect to either of the tools. There was also no statistically significant difference in the relevance of the retrieved pages as measured by the reciprocal answer ranks. Thus, *we were not able to conclusively answer research question Q1* (improvement at the task level) and proceeded to the second phase, which promised more statistical power due to a more efficient design.

Relevance Evaluation Phase
This phase compared only the relevance of the pages returned by the tools. We started the relevance evaluation phase while still waiting to hear from 2 remaining users from the previous phase. We decided that the number of questions obtained from the first 7 users was appropriate for this phase. Since we needed the user queries from the previous phase for the comparison of the results, we limited this phase to only the questions that the users have attempted with Google, thus, we have performed the comparison on 21 questions, which we re-run in a batch mode through both tools. The questions were randomly assigned to the same 9 volunteers as in the previous phase. Using same people was a confounding factor, that is why we minimized its impact by assigning question in such a way that no user was judging the relevance of the query results to his/her own questions.

The judges were presented with the questions and the retrieved results (limited to top 10) in the same format as during the previous phase. This time, the users acted as "blind" judges, since they did not know which tool produced the results. The retrieval results from both tools were assigned to the judges at random but while observing the following constraints: 1) each question was assigned to at least two judges (each assignment produced by a different tool). 2) No question was assigned to more than four judges (2 from each tool). 3) No judge was assigned a question that he/she contributed during the previous phase. This assignment allowed avoiding "familiarity" effects and also distributing questions more uniformly than with completely random assignment. The instructions to locate the first answer page were essentially the same as in the previous phase. The judges were not forced to pick any relevant page, thus were allowed to state that "nothing suitable was found."

Thus, our operatioanalized hypothesis was the following:

$H1_o$: *QA tool produces same reciprocal rank as Google*. The alternative hypothesis $H1_a$ was that QA performed better.

Table 2 (posted at http://www.public.asu.edu/~droussi/IEEE2005-short-tables.html)shows the evaluation results for each question. Since two judges did not return their surveys by the time of our statistical analysis, there are some missing values in the table. We had to reject both $H1_o$ at the levels of confidence $\alpha < 0.1$ *(p-value = 0.08)*, thus empirically confirming that the overall relevance of the returned pages was better with QA tool and answering positively our second research question. The relative improvement was quite substantial: by using Question Answering technology, the average reciprocal ranks were increased by up to 25%, which we believe is a practically important result.

Qualitative Observations

We have observed that the following major categories of web pages were indicated by our users as providing the answers to their questions (in the approximate order by their frequency of occurrence), and thus potentially could contain malevolent content:

1) Message boards, newsgroups or discussion forums. Those sites typically allow unrestricted posting of messages, anonymous or using an alias. Very often, the answer found is a post, titled as a response to another post, e.g. "Re: I wanna crack hotmail passwords."

2) Stand alone web pages that may be considered "hacking" tutorials, guidance, advice etc. They frequently have "How to" in the title or headings. Many of the web sites hosting them have international (non US) domain names.

3) Technical documentation from software vendors. Those pages would be typically posted for cyber security experts or just general computer users, and serve completely legitimate purpose of educating on the issues of security, e.g. for the purpose of avoiding fraud or security breaches. However, many potential intruders can get educated there as well.

To preserve the anonymity of the owners of the websites and the authors of the posts we decided not to mention any specific URLs or domain names in our paper.

User	Question Number	Reciprocal Answer Rank QA	Reciprocal Answer Rank Google
1	1	0.25	0.26
	3		0.2
	5	0.16	1
2	1	0.6	1
	3	0.75	
	5	0.25	
3	1	1	0
	3		
	5	0.25	0
4	1		0.5
	3		0.05
	5		
5	1	0.14	0.05
	3	0.75	
	5	0.25	0.33
6	1	0.25	0.25
	3		0.42
	5	1	0.5
7	1	0.42	
	3	0.5	
	5		0.5
Mean:		0.470	0.364
Mean Standard Deviation		0.024	0.028

Table 2. Blind relevance judgments

ACKNOWLEDGEMENTS

CONCLUSIONS AND FUTURE RESEARCH

We have established that pattern based Question Answering (QA) technology is more effective at locating web pages that may provide answers to the set of indicative questions (such as "How do I crack passwords?", "How do I steal a credit card number?" etc.) and, by this, delivering potentially malevolent content. Thus, QA can be used as a substitute or in addition to traditional keyword searching by law enforcement, public watchdogs and researchers whose responsibilities include locating and monitoring online content. As a by product, we compiled a collection of "indicative" questions to which potential cyber violators may be seeking answers. Those questions can be used in follow up batch mode tests.

Our direct implications are to the law enforcement officers and to the law-enforcement IT systems designers. Our results suggest that in addition to the popular keyword search it pays off to invest in the tools that are based on Question Answering technologies, specifically those similar to our pattern based probabilistic approach. QA approach to locating malevolent content is also more intuitive and natural than the traditional keyword approach since the interaction is based on natural language questions, which are easier to form than keyword queries, especially for novices in searching technology. Even the law enforcement professional may be skilled at using advanced search engine features, lowering the expectation of expertise may be beneficial for less adept searchers such as assisting staff, trainees and public volunteers.

We focused our efforts on the technical aspect of the problem and deliberately have not considered any ethical or political aspects of the problem, for example the legal and political feasibility of removing malevolent content from the web or prosecuting the authors. The literature reviewed in our introduction suggests that such actions are possible and carried out in practice.

REFERENCES

1. Brin, S. and Page, L. (1998) The Anatomy of a Large Scale Hypertextual Web Search Engine, Stanford technical report, Stanford Database Group Publication Server [WWW] http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en &doc=1998-8&format=pdf&compression= (February, 2005)

2. Claburn, T. (2005) Search For Tomorrow, *Information Week*, March 28, 2005.

3. Dumais, S., Banko, M., Brill, E., Lin, J. and Ng, A. (2002) Web Question Answering: Is More Always Better? In *Proceedings of the ACM Conference on Information Retrieval,* ACM.

4. Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Girju, R., Rus, V. and Morarescu, P. (2000) Falcon: Boosting knowledge for answer engines. In Voorhes, E.M., and Harman, D.K. *Proceedings of the Ninth Text REtrievalConference (TREC 9)*, November 13-16, Gaithersburg, Maryand, USA, NIST, 479–488.

5. Jacques, R. (2004) 28 Arrested in Global Web Fraud Sting. Personal Computer World, October 2004 [WWW] http://www.pcw.co.uk/news/1159056 (February, 2004)

6. Lempert, R. J., Popper, S. W., Bankes, S. C. (2003) Shaping the next one hundred years: new methods for quantitative, long-term policy analysis, RAND, Santa Monica, CA.

7. Leouski, A., & Allan, J. (1998) Visual Interactions with a Multidimensional Ranked List. In *Proceedings of the Twenty First Annual International ACM Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 353-354.

8. Lyman, P., and Varian, H.R. (2000) How Much Information?, School of Information Management and Systems, at the University of California at Berkeley [WWW] http://www.sims.berkeley.edu/research/projects/how-much-info/ (February, 2005)

9. National Science Foundation. (2003) NSF Announces $30 Million Program in "Cyber Trust." NSF Web site [WWW] http://www.nsf.gov/od/lpa/news/03/pr03133.htm (February, 2004)

10. Radev, D. R., Libner, K., & Fan, W. (2001) Getting answers to natural language queries on the web. *Journal of the American Society for Information Science and Technology (JASIST)*, 53, 5, 359-364.

11. Radev, D., Fan, W., Qi, H., Wu, H., Grewal, A., (2002) Probabilistic question answering on the web. In Lassner, D. (Ed.) *Proceedings of the 11th WWW conference*, May 7-11, Honolulu, Hawaii, USA, Association for Computing Machinery.

12. Radev, D., Fan, W., Qi, H., Wu, H., Grewal, A., (2004) Probabilistic question answering on the web, accepted for publication. Journal of the American Society for Information Science and Technology.

13. Reid, E., Qin, J., Chung, W., Xu, J., Zhou, Y., Schumaker, R., Sageman, M. and Chen, H.(2004) Terrorism Knowledge Discovery Project: A Knowledge Discovery Approach to Addressing the Threats of Terrorism, in *Proceedings of the Second Symposium on Intelligence and Security Informatics*, June 10-11, Tucson, AZ, USA, 125-145.

14. Roussinov, D. and Chen, H. (2001) Information navigation on the web by clustering and summarizing query results. *Information Processing and Management*, 37, 6, 789-816.

15. Roussinov, D. and Robles-Flores, J.A. (2004) Web Question Answering: Technology and Business Applications, in *Proceedings of the Tenth Americas Conference on Information Systems*, August 6-8, New York, USA, 3248-3254.

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

17. Shneiderman, B., Byrd, D., and Croft, W.B. (1998) Sorting out searching: A user- interface framework for text searches. *Communications of the ACM*, 41, 95-98.

18. Soubbotin, M., and Soubbotin, S. (2002) Use of patterns for detection of likely answer strings: A systematic approach. In Voorhees, E.M., and Buckland, L.P. (Eds.) *Proceeding of the Eleventh Text REtrieval Confeence (TREC-11)*, November 19-22, Gaithersburg, Maryland, USA, NIST.

19. Swartz, J. (2004) Hackers hijack federal computers, USA Today, [WWW] http://www.usatoday.com/tech/news/computersecurity/2004-08-30-cyber-crime_x.htm (February, 2005)

20. Turetken, O., Sharda, R. (2004) Development Of A Fisheye-Based Information Search Processing Aid (FISPA) For Managing Information Overload In The Web Environment, *Decision Support Systems*, 37, 3, 415-434.

21. Verton, D. and Verton, D. (2003) Black Ice: The Invisible Threat of Cyber-Terrorism, McGraw-Hill Osborne Media, Emeryville.

22. Voorhees, E. and Buckland, L., Eds. (2003) *Proceedings of the Twelfth Text REtrieval Conference TREC*, November 18-21, Gaithersburg, Maryland, USA, NIST.

23. Weimann, G. (2004) How Modern Terrorism Uses the Internet. SPECIAL REPORT 116, United States Institute of Peace [WWW] http://www.usip.org/pubs/specialreports/sr116.html (February, 2005).

24. Wells, H.G., (1895) The Time Machine. Tor Books; Reissue edition (December 1, 1995), New York.