A Technique to Resolve Contradictory Answers

Catherine Blake

Department of Information and Computer Science
University of California, Irvine
cblake@ics.uci.edu

Abstract

With the ever-increasing amount of textual information available, it is becoming increasingly unlikely that a single document will provide the answer to a question. A more likely scenario is that multiple documents will each propose answers. Current question-answering systems rarely resolve redundant or contradictory information when answering questions. This is a problem in the biomedical community where multiple documents often attempt to answer a question. The multi-layer question answering approach proposed in this position paper was inspired by a study of scientists in public health and medicine as they used the medical literature to answer research questions. The proposed approach, called Information Synthesis (IS), takes advantage of existing question answering techniques to extract facts from biomedical literature then incorporates formal meta-analysis techniques to unify contradictory findings. This approach is based on the premise that the person posing the question has knowledge of the information required and integration methods required to answer the question, which is supported by our own experience and with the user group studied.

Introduction

The amount of information available in electronic form continues to increase at an alarming rate. Question answering systems are well suited to users who have a specific information need. An accurate question answering system reduces information overload, by providing not only the document that contains the information of interest, but the actual answer to the question from within a document. The implicit assumption in this scenario, and thus in the question answering evaluations, is that at least one document within the corpus reports the answer to a question. In the case where multiple documents propose an answer (or multiple answers are reported in a single document), a question answering system typically identifies from the set of possible answers, a single 'correct' answer. For example, the AskMSR system identifies the most likely answer based on answer frequency (Banko, Brill et al. 2002).

Although the current question-answering paradigm suggests a single correct answer, redundant and contradictory answers are critically important to scientists in the biomedical community. Consider the question *What is the relationship between alcohol consumption breast cancer risk?* A recent analysis (Ellison, Zhang et al. 2001) found more than seventy articles that each proposed an answer to the alcohol-breast cancer question. Meta-analysis (the approach used by Ellison and his colleagues) offers one approach to unify the redundant and contradictory answers proposed by each study. Originally developed in the agricultural community, meta-analysis has gained popularity by scientists in medicine and public health as an effective technique to resolve different findings from multiple studies. A detailed description of meta-analytic techniques can be found in (Ingelfinger, Mosteller et al. 1994).

The Need for Information Synthesis

Consider again the question *What is the relationship between alcohol consumption breast cancer risk?* One way to answer this question is to decompose it into a set of questions that can be answered using the document corpus. For example, *Do people with breast cancer consume significantly more alcohol than people in a similar population who do not have breast cancer?* (see figure 1) Comparing alcohol consumption rates of people with breast cancer with consumption rates of people without breast cancer would provide an answer to the decomposed question (figure 1, level 3). To answer this question a scientist would require an answer to the questions *How much alcohol do people with breast cancer consume* and *How much alcohol do people without breast cancer consume* (figure 1, level 2). Each breast cancer study that reported alcohol consumption would provide an answer to the former question. Some studies that report alcohol consumption would also report alcohol consumption for subjects without breast cancer. For studies that only report the rate of breast cancer subjects, an external source could be used to estimate a comparison population. For example, the age, gender, time-frame and geographical location of a study could be used to identify a comparison population from the Behavioral Risk Factors Surveillance (BRFSS) system (www.cdc.gov). The BRFSS provides the results of an annual survey that explores health behaviors of US residents, such as alcohol and tobacco consumption. A significantly higher rate of alcohol consumption in breast cancer subjects than in the comparison group (after controlling for study size), would provide strong evidence that alcohol consumption increases breast cancer risk.

The shaded area in figure 1 indicates questions asked within the context of an individual study. The answers to questions such as the location and time-frame of the study are used to identify a comparison group from the BRFSS. Only the first question identified in level 2 is addressed in figure 1.

information synthesis is appropriate for the class of questions that users can reformulate as a question with a quantitative answer. For complex domains such as biomedicine, multiple answers, gained directly from a document, or via information synthesis, could be the only accurate answer to a question.

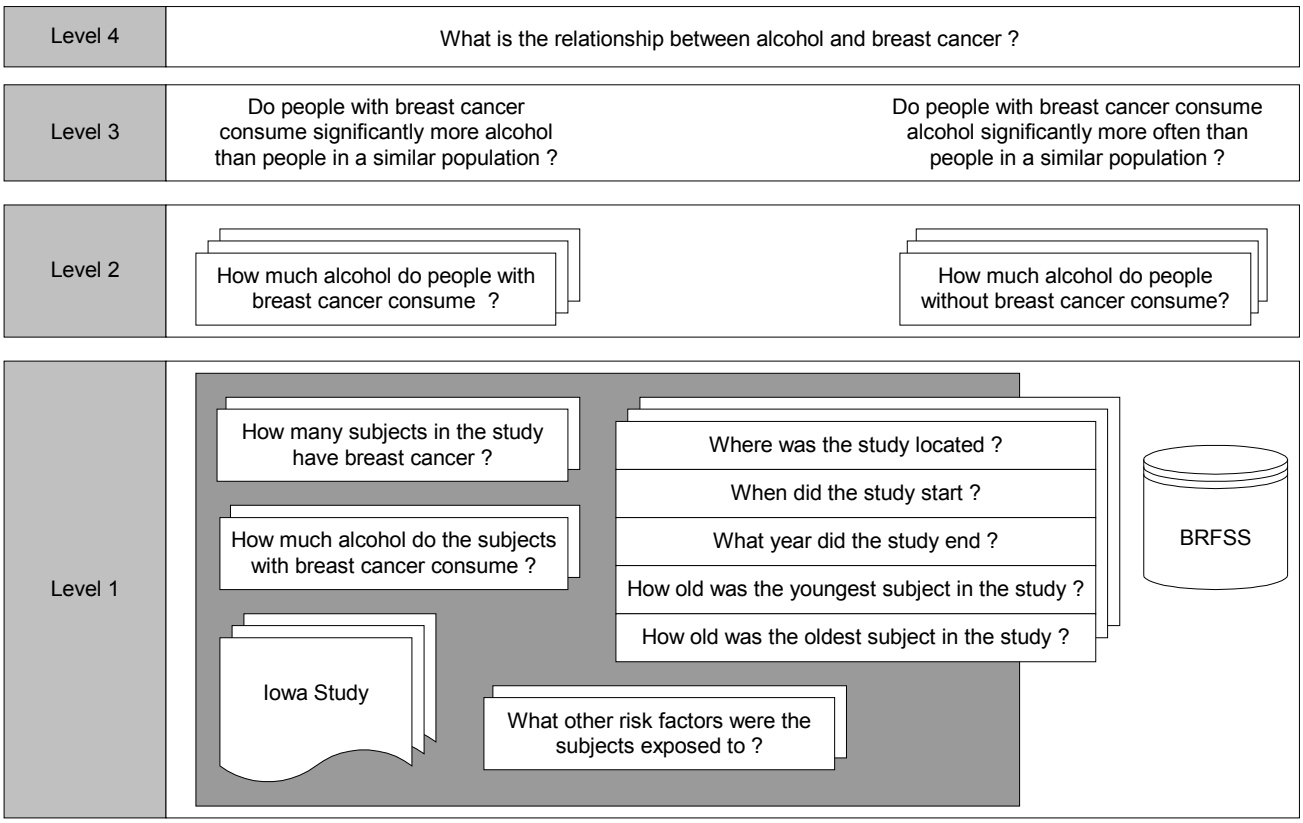

Figure 1 - Information synthesis incorporates multi-layered question answering to address the situation of differing answers to a question.

Discussion

The query reformulation described in the previous section offers provides a conceptual framework to resolve contradictory answers for the next generation of question answering systems. A study that used this model found that the current smoker rate of men in impotent studies was significantly higher than the rate of current smokers reported in the BRFSS who were similar with respect to age, geographical location, time-frame and gender (Tengs and Osgood 2001). Although the authors used manual techniques to identify and extract information from the 1008 articles considered, there are clearly opportunities to semi-automate phases of this process using question-answering technology.

Unlike summarization techniques that tackle the difficult task of generating an answer in natural language, In addition to unifying different answers to a question, the information synthesis process can reduce the effect of publication bias. Numerous studies have shown that epidemiological investigations that show a significant relationship between a risk factor and a disease are more likely to be published than studies that do not show such a relationship (Easterbrook, Berlin et al. 1991). When these studies are then pooled in meta-analysis, a biased impression of the true impact of certain risk factors is codified (Begg and Berlin 1988). For example, seventy-one of the seventy-two articles in the previously mentioned alcohol-breast cancer meta-analysis contained the word alcohol or a synonym in the title or abstract. Thus, the primary purpose of each study considered was alcohol consumption and subject to this form of publication bias. Using an external data source enables studies that report alcohol consumption as secondary information to be included in the analysis. For example (Zheng, Deitz et al.

1999) and (Helzlsouer, Alberg et al. 1999) report alcohol consumption, however these studies were not considered in the alcohol-breast cancer meta-analysis. A combination of this unintentional exclusion and publication bias may be responsible for the elevated risk reported in analysis.

In our study of scientists as they used biomedical literature to answer research questions (Blake and Pratt In Press), we found that similar information items were used, despite very different topics. The medical scientists we studied were focused on the question *What is the reliability of spinal palpatory procedures?* The group in public health set out to explore *What is the relationship between smoking and impotence.* Both studies required information items related to the population studied (such the age of subjects and geographical location) and the study design (such as the number of participants). This regularity suggests that levels within the information synthesis model can be re-used at higher levels.

Clearly, information synthesis will not offer a substitute to clinical trials or cohort studies, however this approach could accelerate the time required by scientists to address a specific question using on currently reported findings. The proposed information synthesis process assumes that users know what information is required and how to integrate the information.

Although motivated by biomedical questions, information synthesis could also be used to answer questions in other domains, such as in the investment question shown in figure 2.

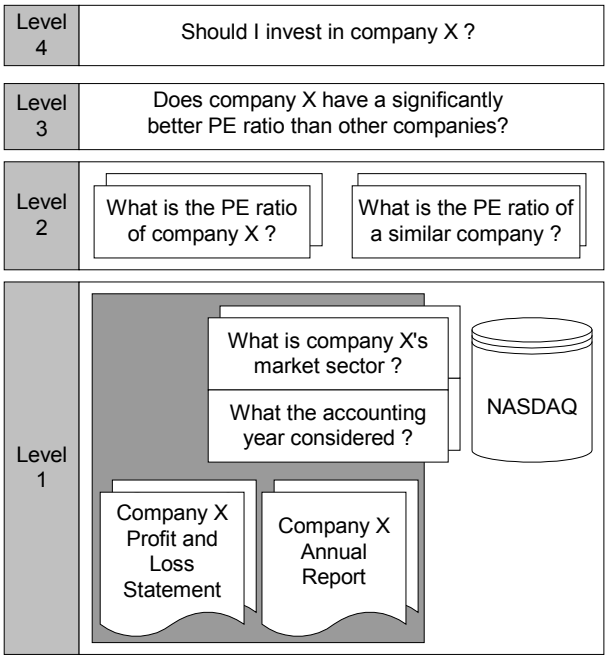

Figure 2 – The information synthesis process used to answer an investment question.

Current and Future Work

This position paper describes the framework we are using to develop a semi-automated approach to answer the question *What is the relationship between smoking breast cancer risk?* We have not automated the conceptual reformulation of a question, however we have developed a pilot system that extracts level 1 answers such as those shown in figure 1 (Blake, Pratt et al. In Press). After manual verification of the information extracted, the current system conducts a meta-analysis as outlined in figure 1. Working with an expert in breast cancer to

References

Banko, M., E. Brill, et al. (2002). AskMSR: Question Answering Using the Worldwide Web. Mining Answers from Texts and Knowledge Bases, AAAI Spring Symposium, Technical Report SS-02-06. S. M. Harabagiu and V. Chaudhri. Stanford, AAAI Press.

Begg, C. and J. Berlin (1988). "Publication bias: a problem in interpreting medical data." Journal of the Royal Statistical Society(151): 419-63.

Blake, C. and W. Pratt (In Press). Collaborative Information Synthesis. American Society of Information Science and Technology, Philadelphia.

Blake, C., W. Pratt, et al. (In Press). Automated Information Extraction and Analysis for Information Synthesis. AMIA, San Antonio.

Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services.

Easterbrook, P. J., J. A. Berlin, et al. (1991). "Publication bias in research." Lancet **337**: 867-72.

Ellison, R. C., Y. Zhang, et al. (2001). "Exploring the Relation of Alcohol Consumption to Risk of Breast Cancer." American Journal of Epidemiology **154**(8): 740-747.

Helzlsouer, K. J., A. J. Alberg, et al. (1999). "Serum concentrations of organochlorine compounds and the subsequent development of breast cancer." Cancer Epidemiology, Biomarkers & Prevention **8**(6): 525-32.

Ingelfinger, J. A., F. Mosteller, et al. (1994). Biostatistics in clinical medicine, McGraw-Hill Inc.

Tengs, T. and N. D. Osgood (2001). "The link between smoking and Impotence : Two Decades of Evidence." Preventive Medicine **32**(6): 447-452.

Zheng, W., A. C. Deitz, et al. (1999). "N-Acetyltransferase 1 Genetic Polymorphism, Cigarette Smoking, Well-done Meat Intake, and Breast Cancer Risk." Cancer Epidemiology, Biomarkers & Prevention **8**(3): 233-9.