Reasoning about Knowledge from the Web - (Extended Abstract)

Gjergji Kasneci*

Hasso-Plattner-Institute,
Prof.-Dr.-Helmert-Str. 2-3,
14482 Potsdam, Germany
Gjergji.Kasneci@hpi.uni-potsdam.de

In the presence of a vast amount of user generated content evolving around entities such as people, locations, products, events, etc., it seems that document-oriented retrieval is rather old-fashioned. Imagine an HIV-relevant search task with the goal of finding drugs that may interfere with HIV protease inhibitors. Retrieving an exhaustive list of explicit results (i.e., drugs that may interfere with HIV protease inhibitors) can be crucial for people suffering from HIV, whose health depends on the unmediated effect of protease inhibitors. Moreover it might be desirable to have the drugs in the result list ranked by their probability of interfering with protease inhibitors. In order to automatically retrieve such an exhaustive list of ranked answers, there are two subtasks that have to be addressed: (1) knowledge about drugs that stand in an interference relationship to protease inhibitors needs to be extracted from various web pages and appropriately combined, (2) the drugs need to be ranked by interference probabilities. Neither of these tasks can be addressed by state-of-the-art search engines. Expecting the user to manually inspect retrieved documents to construct an exhaustive list of answers is simply unrealistic. As a matter of fact, major players in the search engine industry have recognized these issues and are attempting to shift the focus towards knowledge retrieval. For example, in 2010, Google acquired Metaweb, the company behind Freebase, one of the largest knowledge bases with explicit facts about real-world entities. In 2011, Google's search group was restructured and renamed into "knowledge group" [6]. Another example is Microsoft's Bing, which has undergone similar changes in recent years. By the end of 2009 Bing was returning Wolfram Alpha results to entity-related and scholarly queries [8], and by the end 2010 Bing announced the new "health search experience" with the focus "on further enabling people to get relevant information and make better decisions" [7].

Some years earlier, two outstanding academic efforts [3, 4] proved the concept of knowledge base construction with facts extracted from semi-structured information sources in Web 2.0 platforms. The information in such sources is typically contributed and curated by many different users, thus reflecting the "Wisdom of the Crowds". The most well-known example in this realm is Wikipedia, which provides infoboxes, categories, and other kinds of tabular information about the

* I am grateful to Thore Graepel and Jurgen Van Gael for many insightful comments and discussions on this topic.

entities described by the articles. Such assets mitigate the need of Natural Language Processing and Statistical Learning techniques for information extraction and allow instead the adoption of much simpler techniques such as regular expressions, lexicons, and pattern matching algorithms. Although the knowledge bases derived by the latter extraction techniques from Web 2.0 sources have a relatively high coverage and quality, much of the knowledge they contain is inherently uncertain. Quantifying this uncertainty is a major concern, which, to a large extent, has been ignored by the semantic web community.

To address the above concern, we propose a probabilistic knowledge representation model that quantifies uncertainty by exploiting user feedback on the truth values of statements in the knowledge base [1]. In this model the truth value of each statement is represented by a binary random variable and the logical interdependencies between statements, such as transitivity (e.g., if Potsdam is located in Brandenburg and Brandenburg is located in Germany then Potsdam must be located in Germany too), are represented as a Bayesian network connecting the binary random variables. In order to capture feedback on the statements we introduce binary random variables standing for the user feedback and continuous variables representing the reliability of users. The latter feedback components are directly connected with the random variables representing the truth values of statements. Note that the user feedback initiates and enables updates on the beliefs of the variables in the network. A sample subgraph from the Bayesian network is presented in Figure 1.

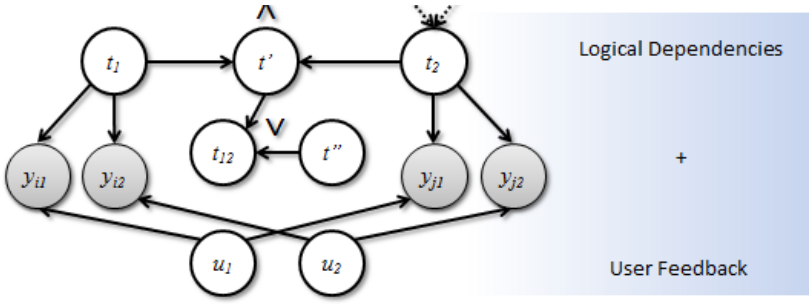

Fig. 1. Belief updates through feedback propagation.

Figure 1 illustrates how the model enables joint reasoning on the reliability of two users (represented by the variables u_1, u_2) and the truth values of two statements (represented by the variables t_1, t_2) on which they provide feedback. Additionally, the truth value of a statement t_{12} that is deduced from the former statements is learned through the probabilistic conjunction (represented by t') of t_1, t_2 and its disjunction with a variable t'', which accounts for incomplete knowledge, i.e., any deductions which might not be captured by the knowledge base.

As the reliabilities of users vary across knowledge domains, we also propose an extension of the above model, in which the user expertise is measured by means of a collaborative-filtering-style probabilistic model [2]. More specifically, in this model, users and statements are represented by feature vector variables (e.g., user features are: id, gender, country, etc., and statement features: id, topic, relationship type, etc.), which are mapped into a lower dimensional latent space where the similarity between users and statements can be measured.

For the probabilistic inference in the above models we have used Expectation Propagation as implemented by Infer.NET [5].

In experiments with a subset of YAGO statements [4] and feedback collected from Amazon Mechanical Turk[1], both models described above turn out to be far more accurate than a model that aggregates feedback based on majority voting[2]. Furthermore, a high prediction accuracy can be already achieved with relatively sparse feedback, thus avoiding unrealistic effort on the users' side. Finally, the model that captures the expertise of users excels in both accuracy and reduction of the amount of feedback needed.

Based on a distributed computing framework, we implemented an efficient, large-scale version of the above models, which can handle knowledge bases with hundreds of millions of statements. The result was presented at Microsoft's largest internal research and technology fair, TechFest 2011.

References

1. Kasneci, G., Gael, J. V., Herbrich, R., Graepel, T.: Bayesian Knowledge Corroboration with Logical Rules and User Feedback. In: Machine Learning and Knowledge Discovery in Databases (ECML 2010), pp. 1–18. Springer (2010)
2. Kasneci, G., Gael, J. V., Stern, D. H., Graepel, T.: CoBayes: Bayesian Knowledge Corroboration with Assessors of Unknown Areas of Expertise. In: International Conference on Web Search and Web Data Mining (WSDM 2011) pp. 465–474. ACM (2011)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. G.: DBpedia: A Nucleus for a Web of Open Data. In: International Semantic Web Conference (ISWC/ASWC 2007). pp. 722-735. Springer (2007)
4. Suchanek, F. M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. J. Web Sem. 6(3): 203–217 (2008)
5. Infer.NET : `http://research.microsoft.com/en-us/um/cambridge/projects/infernet/`
6. Google Dissolves Search Group: `http://techcrunch.com/2011/05/03/google-dissolves-search-group-internally-now-called-knowledge`
7. Bringing Knowledge into Health: `http://www.bing.com/community/site_blogs/b/search/archive/2010/01/12/bringing-knowledge-into-health-search.aspx`
8. Bing, Wolfram Alpha Agree on Licensing Deal: `http://www.zdnet.com/news/bing-wolfram-alpha-agree-on-licensing-deal/333870`

[1] For an exact description of the experimental setting, we refer the reader to [1].

[2] In the majority-based aggregation of feedback the truth value of a statement is estimated by a majority consensus; that is, a statement is assumed to be true if the majority thinks that it is true.