Automatic Construction of a Semantic, Domain-Independent Knowledge Base

David Urbansky

University of Technology Dresden, Department of Computer Science
david.urbansky@tu-dresden.de

Abstract. In this paper, we want to show which difficulties arise when automatically constructing a domain-independent knowledge base from the web. We show possible applications for such a knowledge base to emphasize its importance. Current knowledge bases often use manually-built patterns for extraction and quality assurance which does not scale well. Our contribution to the community will be a technique to automatically assess extracted information to ensure high quality of the information and a method of how the knowledge base can be kept up to date. The research builds upon the existing WebKnox system for Web Knowledge Extraction which is able to extract named entities and facts from the web. This is a position paper.

1 Introduction

The web hosts millions of information pieces and is still growing at a rapid pace. No single human can have an overview of all web pages and the information they provide, thus, the trend towards a machine "understandable" web has been proposed in the semantic web initiative [3]. If machines can read, "understand" (disambiguate) and aggregate information pieces from many different sources, the human can consume the desired information much faster. In different knowledge domains the approaches for retrieving and extracting knowledge can vary. This requires more human effort in configuring the extraction system. We want to minimize the human effort and therefore need a generic system that automatically builds a structured knowledge base from the information that is available on the web.

Such a knowledge base has many possible applications. We pick two application domains which seem most important for our research:

1. **Entity Dictionary for Information Enrichment.** Imagine the possible scenario: A user reads the name of a movie that he does not know on a web page, he or she might leave the page and search for the movie on a search engine to learn more about it. If we have a large knowledge base with semantic information about entities (such as movies or persons), we can recognize and enrich them for a human user. In the described scenario, we could find additional information such as the director or release date of the mentioned movie and provide the user with this information directly on

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 800–804, 2009.

the web page. Since the information from the knowledge base is semantic, we could also disambiguate entity names, that is, we would provide information about the person "Queen Elizabeth II" on a web page about the English queen and information about the ship "Queen Elizabeth II" on a web page about travel and cruise ships.

2. **Repository for the Semantic Web.** In the vision of the semantic web [3], the information on the web is becoming more machine readable. This allows many new applications, that the human end user can benefit from. Today, many domain ontologies have been created and connected (DBPedia, MusicBrainz etc.) in order to allow machines to reason over that information and gain new insights for the human. The desired knowledge base will be centralized, connected to existing ontologies and thus, be a part of open information for the semantic web. The ontology will be represented in RDF(S) and OWL and users can query the knowledge base using SPARQL.

The remainder for this paper is structured as follows. First, we review systems for information extraction, aggregation and knowledge base building. Second, we describe shortcomings of these systems, third, we pose research questions and explain how we want to address these shortcomings in the future, fourth, we describe the current state of WebKnox (Web Knowledge eXtraction), a system for information extraction from the web, and finally we conclude the paper with an outlook on the working plan.

2 Related Work

In this section, we review several significant domain-independent information extraction systems and knowledge bases.

KnowItAll [5], **TextRunner** [2], **GRAZER** [11] and **Set Expander for Any Language** [9] are domain-independent, information extraction systems that automatically extract entities and facts from the web. All of these systems are mainly redundancy-based, that is, that they rely on the assumption that a fact or entity occurs many times on the web. For ensuring a high precision of the extracts, assessment mechanisms such as statistical methods or graph algorithms are used.

Freebase[1], **DBpedia**[1], **YAGO**[6], **Wolfram Alpha**[2] and **True Knowledge**[3] are semantic knowledge bases. In contrast to the first group of systems, these knowledge bases were constructed using domain- or website specific extraction algorithms. Freebase, Wolfram Alpha and TrueKnowledge rely on human user input to keep information up to date. The other knowledge bases are kept up to date by regularly re-indexing their information.

All of the mentioned systems and knowledge bases have insufficiencies in one or more of the following areas that we want to address with WebKnox. First

[1] `http://www.freebase.com`, accessed 16/08/2009
[2] `http://www.wolframalpha.com`, accessed 16/08/2009
[3] `http://www.trueknowledge.com`, accessed 16/08/2009

of all, we can not only rely on a few sources and human editors if we want to have a broad knowledge base, that is, we need to treat all web pages as possible sources of information. Second, we need to store the extracted information semantically, third, we need to automatically assess extracted information ensuring high quality of the knowledge base and finally we need to find efficient, automatic techniques to keep the knowledge base up to date. In the following section, we pose the research questions that address these requirements.

3 Research Questions and Approaches

How can we ensure high precision of the extracted information? Since we want as little human involvement as possible assessing the extractions in the knowledge base, we must ensure that uncertain extractions are automatically ranked by their confidence, that is, we need to find and evaluate "trust" measures for sources and extractions. Several approaches have been used in current state-of-the-art systems, such as URNS [4], Pointwise-Mutual-Information (PMI) [5], Random Graph Walk [9] and using search engine rankings and duplication information [10]. In [8], we have proposed a self-supervised machine learning algorithm for scoring fact extractions. We now want to explore how good similar machine learning classification algorithms perform on extracted entities. For example, if the movie entity "The Dark Knight" was extracted many times, using several of the entity extraction techniques, we should have a higher confidence in its correctness as if it was only extracted once from one single source. Also textual features such as capitalization of the names can help classifying entities.

How can we keep the knowledge base up to date? Unlike humans, an automated system can read millions of documents each day. That enables the system to find and extract new knowledge faster. We want to investigate into automatically reading news feeds (blogs, forums etc.) to extract new knowledge. For example, letting WebKnox read thousands of news feeds about movies and cinema, we could extract information about the upcoming movie "Iron Man 2" very quickly when a news item states "[...]has signed for the upcoming movie Iron Man 2[...]".

Where and how can we find new entities? Currently, our system uses three retrieval and extraction techniques to find and extract entities for known concepts from the web [7]. So far, we evaluated the entity extraction only on popular concepts such as mobile phones or countries. To find entities for more obscure concepts such as cookies, perfumes or pencils we need to find other mechanisms. One approach in this direction is to analyze user queries for possible unknown entities. For example, if we know the concept "cookie" and the query "calories Banana Split Creme Oreo" is received, we could try to find out that "calories" is an attribute of the "cookie" concept and the rest of the query is an entity. Also, we will investigate whether domain specific databases from the Deep Web can help finding new entities.

4 Current State of WebKnox

WebKnox is divided into two main extraction processes as shown in Figure 1. First, the entity extraction process gets a set of predefined concept names as input (for example "Movie", "Mobile Phone" or "Country") and then queries a multi-purpose search engine such as Google to retrieve pages with possible entity occurrences. The extracted entities are then written into the knowledge base. After the knowledge contains some entities, the fact extraction process reads those entity names and also gets predefined information about the attributes that are searched for the entity's concept. For example, the fact extraction process reads the movie entity "The Dark Knight" from the knowledge base and the attributes "runtime", "director" and "release date" from the knowledge ontology. The process then queries a search engine again, tries to extract facts, assesses the extractions and writes the results back to the knowledge base. More details can be found in [8] and [7].

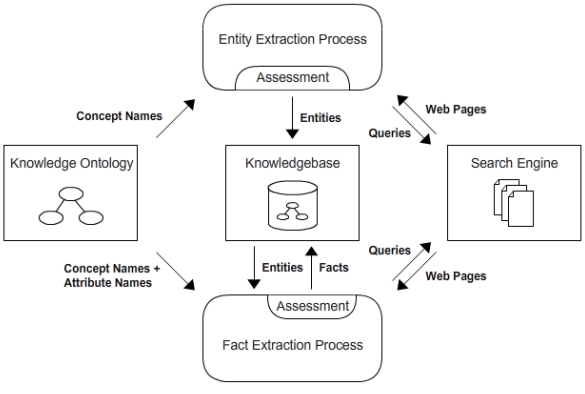

Fig. 1. Overview of extraction flow

5 Conclusion and Work Plan

In this paper we have shown existing approaches to extract information from the web and build semantic knowledge bases. We have described the shortcomings of these systems and how we are planning to address them. Our future work is structured as follows.

1. **Create Use Cases** First, we will describe use cases in each of the application domains for a semantic knowledge base. The use cases will help to understand the requirements for such a knowledge base in more detail.
2. **Research on Related Work** Second, we are going to find state-of-the-art approaches for knowledge base updating and extraction assessment. We will also need to implement some of the most relevant algorithms (PMI, URNS, random graph walk) to have a baseline for our own approach.

3. **Design and Experimentation** In the third phase, we will create verifiable hypotheses that lead the design and experimentation process for our own solutions. This phase will start early, as practical problems can be recognized quickly.
4. **Evaluation** In the last phase, we will create a real life testing set for our assessment approach and compare it to the identified state-of-the-art algorithms. Furthermore, we will evaluate how quickly the knowledge base is updated with correct information in different domains.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A nucleus for a web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
2. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open Information Extraction from the Web. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 2670–2676 (2007)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 28–37 (2001)
4. Downey, D., Etzioni, O., Soderland, S.: A Probabilistic Model of Redundancy in Information Extraction. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 1034–1041. Professional Book Center (2005)
5. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the Web: An experimental study. Artificial Intelligence 165(1), 91–134 (2005)
6. Kasneci, G., Ramanath, M., Suchanek, F.M., Weikum, G.: The YAGO-NAGA approach to knowledge discovery. SIGMOD Record 37(4), 41–47 (2008)
7. Urbansky, D., Feldmann, M., Thom, J.A., Schill, A.: Entity Extraction from the Web with WebKnox. In: Proceedings of the Sixth Atlantic Web Intelligence Conference (to appear, 2009)
8. Urbansky, D., Thom, J.A., Feldmann, M.: WebKnox: Web Knowledge Extraction. In: Proceedings of the Thirteenth Australasian Document Computing Symposium, pp. 27–34 (2008)
9. Wang, R.C., Cohen, W.W.: Language-Independent Set Expansion of Named Entities Using the Web. In: The 2007 IEEE International Conference on Data Mining, pp. 342–350 (2007)
10. Wu, M., Marian, A.: Corroborating Answers from Multiple Web Sources. In: Proceedings of the 10th International Workshop on Web and Databases (WebDB 2007) (2007)
11. Zhao, S., Betz, J.: Corroborate and Learn Facts from the Web. In: KDD 2007: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge discovery and data mining, pp. 995–1003. ACM, New York (2007)