Future Generation Computer Systems 31 (2014) 111–119

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Guided curation of semistructured data in collaboratively-built knowledge bases

Wolfgang Gassler *, Eva Zangerle, Günther Specht

Databases and Information Systems, Institute of Computer Science, University of Innsbruck, Technikerstrasse 21A, 6020 Innsbruck, Austria

HIGHLIGHTS

- Avoid proliferation of structures in the knowledge base.
- Semantic refinement by resolving homonyms and avoiding synonyms.
- Exploit user's extensive and valuable knowledge.
- Increase the quantity of information contained in the knowledge base.
- Increase the quality of information in the knowledge base.

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form
3 May 2013
Accepted 17 May 2013
Available online 28 May 2013

Keywords:
Collaboration
Semistructured data
RDF
Recommendations
Schema proliferation
Mixed-initiative

ABSTRACT

The collaborative curation of semistructured knowledge has become a popular paradigm on the web and also within enterprises. In such knowledge bases a common structure of the stored information is crucial for providing efficient and precise search facilities. However, the task of refining, extending and homogenizing knowledge and its structure is very complex. In this article we present two paradigms for the simplification of this task by providing guidance mechanisms to the user. Both paradigms aim at combining the power of automated extraction algorithms with the semantic awareness of human users to accomplish this refinement task.

1. Introduction

Large knowledge bases have always been built in a collaborative fashion to collect and archive the common knowledge of groups. Especially with the enormous growth of the internet and the web 2.0 movement, collaboration has been lifted to a new level—online mass-collaboration. One of the most relevant and popular knowledge bases is Wikipedia, which is also based on this mass-collaboration paradigm and is the most important representative of the wiki-concept. The wiki-concept constitutes one of two traditional major paradigms to store information as it stores information as fulltext without any explicit structure. The second paradigm is the storage in (relational) databases which forces the user to store knowledge or information according to a strict and predefined schema. The advantage of such an approach is that all information has to adhere to the same schema and can be searched and presented in a very uniform and hence efficient way. This paradigm is still used in information systems which are focused on one single domain where the contained items are all structured the same way, e.g., a database of movies and reviews. The big disadvantage of this storage paradigm is the unsatisfactory flexibility regarding new structures or content. Changes to the schema are a very time-consuming, tedious and complex task, as it has to be performed manually and the already stored information has to be adapted to the new structure. Thus, the end-user is fixed to a given schema and cannot insert additional information not matching the given schema. For flexible knowledge bases, especially in the area of mass-collaboration, such a restrictive approach is hardly suited, as it can result in a significant loss of information because of the fact that the user cannot insert all information she might want to. This problem was solved by wikis and was one key of success of wiki-systems. Wiki-systems store content in an entirely unstructured manner and can therefore hold any textual information regardless of its structure. Therefore, users do not need to adapt their knowledge to any predefined

* Corresponding author. Tel.: +43 512 507 96892; fax: +43 512 507 96955.
E-mail addresses: wolfgang.gassler@uibk.ac.at (W. Gassler), eva.zangerle@uibk.ac.at (E. Zangerle), guenther.specht@uibk.ac.at (G. Specht).

0167-739X/$ – see front matter
http://dx.doi.org/10.1016/j.future.2013.05.008

schema and are able to insert all information and knowledge they want to. The shortcoming of this structure-less paradigm is its limited search capability. Consider a complex query such as "Which Austrian cities have more than 10,000 inhabitants and have a female mayor who has a doctoral degree?". It is not possible to answer such a query through full-text search which is provided by most wiki-systems. Weikum et al. [1] observed that modern information systems have to be able to support both structured and unstructured data to combine the advantages of both worlds and be able to answer such questions.

The rest of the article is organized as follows. In Section 2, we describe the characteristics of semistructured data. Section 3 contains a description and motivation of the problem which is tackled in this paper. In Section 4, approaches dealing with the problem of refinement after the data has been inserted are described. Section 5 outlines the main idea behind the Snoopy Concept which is a representative of an alignment and refinement of knowledge and structure during the insertion of data. Section 6 concludes the article and describes future work.

2. Semistructured data

The *semistructured data model* incorporates both the paradigm of structured and unstructured storage. The need for such a new storage paradigm already arose in the 90s [2,3]. Back then it became clear that it would be increasingly important to be able to store mostly unstructured data while at the same time providing efficient and structured querying facilities. With the advent of the World Wide Web, which currently forms the largest unstructured knowledge base, it became obvious that such data cannot be fitted into a predefined schema in order to be able to query it. The application of traditional retrieval and extraction techniques to query such unstructured data reached unsatisfactory results as the formulation of structured and precise queries was not possible due to the lack of structure. Thus, the semistructured data model combines both the structured and the unstructured data model and provides a highly flexible way of storing data as it supports the storage of information in a structured way without the need of specifying a predefined schema.

Throughout the last two decades, various models for semistructured data have been developed, like e.g. [4,5]. Currently, the most popular example of the semistructured data model is RDF (Resource Description Framework, W3C recommendation[1]) [6]. RDF basically models knowledge as triples consisting of a subject, a predicate and an object. The subject (also called the resource) is described by multiple pairs of predicates and according objects. The resource is uniquely identified by a URI (Uniform Resource Identifier[2]). Important facts about the University of Innsbruck within a knowledge base can be stored using triples as e.g. in Listing 1.

```
<http://dbpedia.org/../University_Innsbruck>
  <numberOfStudents><26626>
<http://dbpedia.org/../University_Innsbruck>
  <established><1669>
```

Listing 1: Triples.

These predicate–object pairs – in combination with the article URI itself (the resource, in this case University_Innsbruck) – constitute triples. The subjects, predicates and objects are not restricted in any way and can therefore hold any information while at the same time providing structure due to the triple concept as all objects are given context by the according predicates. The triples are machine-readable and processable and thus provide the base for structured access and complex structured queries. In order to query such semistructured RDF knowledge bases, the standard query language is SPARQL (SPARQL Protocol and RDF Query Language) [7].

As RDF is very flexible and is able to link to even external resources by specifying an external URI as the object of a triple, the interlinking between knowledge bases has become very popular. Sir Tim Berners-Lee has coined the term Linked Open Data (LOD) [8] for such linked and semistructured data.

It is important to note that within semistructured systems, users can arbitrarily choose the predicate used for storing information. This fact is very beneficial as it provides a huge amount of flexibility to the users of the system while at the same time – due to the predicate–object format – still features a certain amount of structure. This fact is crucial in online, mass-collaboration information systems, as there are thousands of different users who come from different social levels, backgrounds and edit information of different domains and contexts.

Beside many other approaches aiming at raising the amount of structural information in wiki-systems, RDF increasingly gained ground in many wiki-systems or plugins [9] in order to lift wikis from unstructured black holes to structured knowledge bases. Even Wikipedia articles already contain semistructured data. The tabular aggregation of the most important facts about an article – so-called infoboxes – which are located on the right hand side of many articles were originally not intended for structured and computer-readable access. However, these are now extracted and used as a base for the most important open semistructured knowledge base DBpedia [10], which consists of more than one billion triples extracted from Wikipedia's huge collaboratively built knowledge base.

3. Problem description

The flexibility of the previously described schemaless semistructured storage in combination with collaborative data curation leads to a massive problem. Every single user has her own view of structuring knowledge and information and uses her own terminology. Furnas et al. [11] already showed in the 80s that two people would spontaneously choose the same word for an object with a probability of less than 20%. This suggests that collaboratively built knowledge based on the semistructured model shows a very high proliferation of structures, schemata and vocabulary. The resulting heterogeneous schema impedes the search facilities as a common schema is essential to answer complex structured queries. For example, a user who searches for the value of *numberOfStudents* cannot find information which was stored using the properties *students*, *numberStudents* or *num_students*. Therefore, especially in collaborative knowledge systems, the creation of a common schema without restricting the domain, type or amount of information is desired. Wikipedia is fighting such heterogeneity by introducing collaboratively created templates and the supervision by the committed community. The task of creating structure in Wikipedia is very demanding, which is shown by Boulain et al. [12]. The authors analysed the edits in Wikipedia and identified that only 35% of all edits within Wikipedia are related to content, whereas all other edits aim at enhancing the structure within the Wikipedia knowledge base. Additionally, Wu and Weld [13] showed that infoboxes which adhere to predefined templates are still divergent and noisy.

Another problem in collaboratively built knowledge bases is the barrier for new users to insert information to knowledge bases. In Wikipedia most of the content is created by a very small group of users. Furthermore, articles have to conform to many policies and other regulations which increase the barrier for contributions by newcomers [14,15]. But not only in public knowledge bases, moreover and especially in enterprise knowledge bases or wikis,

[1] http://www.w3.org/RDF/, last accessed 2012-10-09.
[2] http://www.w3.org/TR/uri-clarification/, last accessed 2012-10-09.