Descriptive Schema: Semantics-based Query Answering

S. D. Lee, Patrick Yee, Thomas Lee, David W. Cheung, Wenjun Yuan

Department of Computer Science, The University of Hong Kong.
{sdlee,kcyee,ytlee,dcheung,wjyuan}@cs.hku.hk

Abstract. We propose the novel concept of "descriptive schema" (DS). Unlike ordinary database schemas, a DS does not restrict the structure of the underlying database. Rather, it is just a probabilistic description of the structure. When answering keyword queries, DS can be used to improve semantics-based query answering and result ranking.

1 Schema: To have or not to have?

Wikipedia is a rich repository of information. However, facilities to exploit the information are still limited. Although typical search WWW search engines such as Google[1] allow users to look for information using keywords, they lack a schema for formulating the queries precisely.

Besides hyperlinks among the Wikipedia pages, many pages have Category tags as well as Infoboxes, which can be exploited to perform more sophisticated searches. For example, the DBpedia community makes use of these tags to build a database of RDF triplets, allowing more expressive and precise queries in the form of SPARQL to be used to retrieve useful information [2].

The above are two extremes of search and query. In the former case, the user can perform a search easily using relevant keywords, without having to learn the schema's lexicon beforehand. In the latter case, a schema can be used to help specify the query more precisely, but it has a non-trivial learning curve. In this paper, we propose the approach of "descriptive schema" to address these shortcomings. We attempt to strike a balance between the ease of use of a schema-less approach and the high accuracy that a schema-based system can bring us.

2 Descriptive Schema

In this paper, we propose a new concept called "Descriptive Schema" (DS). Unlike XSD (XML Schema Definition), DS is not meant to *prescriptively* mandate a structure on the underlying data. We want to retain the flexibility of free format for the pages. Rather, DS, as its name implies, is *descriptive*. It is only a summary of the structure exhibited by the underlying database. It does not define the structure. The data may occasionally violate the DS.

This tolerance to violations marks our biggest innovation, contrasting with existing approaches. Existing approaches to data modelling use "Prescriptive

Schema", which mandates a rigid structure on the underlying data, with little (if any) tolerance to violations.

We model a DS by a set of rules on the underlying data. There are many possible ways to formulate the rules. One example rule is: "90% of the time, a page of class 'Countries' has value for the field 'capital' in the infobox (infobox for countries)". Note that the rules defined in this way are probabilistic, because they are not satisfied all the time. A DS may thus be considered a summary of the patterns occurring in a database, instead of policies imposed on the data.

The task of discovering a DS from a database is a mining task, which is the problem of finding all rules satisfying a the specified syntax and support thresholds, thus following the data mining model in [3].

3 Applications

Since a DS captures semantical information about the underlying data, it enables a semantics-based approach to answering search queries. We can, for instance, use the DS to help us disambiguate the query, enrich the query with semantical information, as well as using the semantical information to rank the search results. Applications of DS include, but are not limited to, the following:

- Keyword Disambiguation
- Query Augmentation
- Result Ranking
- Data Cleansing
- Guidelines for Authors
- Guided Query Building

4 Conclusions

We have proposed the concept of "descriptive schemas", which is a set of rules obeyed by most of the underlying data, with tolerance for violations. Although the primary goal of devising this novel concept was to help answering keyword queries with an accuracy comparable to databases with prescriptive schemas, we have realized that DS can also be useful for other applications. Future works include exploring further potentials of DS, developing a formalism for it, devising efficient algorithms for mining DS, as well as more in-depth studies of the applications mentioned in this paper.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks **30**(1-7) (1998) 107–117
2. Auer, S., Lehmann, J.: What have Innsbruck and Leipzig in common? extracting semantics from Wiki content. In Franconi, E., Kifer, M., May, W., eds.: ESWC. Volume 4519 of Lecture Notes in Computer Science., Springer (2007) 503–517
3. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Min. Knowl. Discov. **1**(3) (1997) 241–258