Key Recommendations for Infoboxes in Wikipedia

Alexander Larcher, Eva Zangerle, Wolfgang Gassler, Günther Specht
Databases and Information Systems, Institute of Computer Science
University of Innsbruck, Austria
{firstname.lastname}@uibk.ac.at

ABSTRACT

Wikipedia infoboxes represent the semistructured part of data inside a Wikipedia article. The creation of infoboxes is based on the use of templates, which provide the user with a predefined amount of keys. These keys are mainly mandatory and have to be specified in the corresponding infobox along with the corresponding values. This imposes significant limits on the user's actions. We analyzed the development of infoboxes from about 50,000 articles. Furthermore, the templates themselves were analyzed, focusing on ambiguity and multiple usage of keys. Based on these results we introduce a prototype for infobox creation called SnoopBox, which is based on a key recommendation system. This system supports the user creating infoboxes and avoids the tricky use of templates. Both the results of the evaluation of this program and the analysis of infoboxes in Wikipedia are presented.

Categories and Subject Descriptors

H.3.5 [**Storage and Retrieval**]: Online Information Services - Web-based services; H.4.m [**Information Systems**]: Miscellaneous

General Terms

Human Factors, Experimentation, Algorithms

Keywords

Recommendations, Wikipedia, Infobox, Semistructured Data

1. INTRODUCTION

Wikipedia celebrates its 10th anniversary in 2011. However, some of the technologies used are no longer up to date. The users of the World Wide Web are accustomed to a certain grade of usability, at least since the concepts of Web 2.0 became the dominant paradigm. The user support system on Wikipedia is not sufficient, particularly when we take into account the method used to insert infoboxes. Wikipedia does not provide any wizard or other similar support to the user. The user himself has to search for an appropriate infobox template and even if he finds one, it is still not easy to work with it. According to Wu and Weld, the creation of Wikipedia infoboxes is an error-prone "copy and paste" process[3]. To investigate the development of templates and their application over the years, we analyzed data from Wikipedia infoboxes. This showed that the actual design of infobox creation is not very convenient. In this paper, we present a new concept aiming at simplifying the creation of infoboxes in order to allow inexperienced users to insert them into a Wikipedia article with greater ease. To achieve this goal, a key recommendation system and a simplified input interface for key-value-pairs is introduced.

The 22nd ACM Hypertext Conference, Posters and Demos.
HT'11, June 6-9, 2011, Eindhoven, The Netherlands.

2. RESULTS OF THE ANALYSIS

We analyzed the development of infoboxes using so-called Wikipedia dumps. These XML-dumps (of up to 5,5 TB) contain the content of all Wikipedia articles of a certain language and are updated periodically. For the first analysis, a dump including all revisions of every article of the English and the German Wikipedia was used. Some representative templates were selected and the articles containing these templates were examined manually. Furthermore, the part of the code concerning the infobox from every revision of one article was extracted using the DBpedia extraction framework[1]. Every line returned by this tool contains a complete key-value-pair of the infobox instance in an article.

The results of this analysis showed that most of the changes in the analyzed articles' infoboxes were updates of lines (86%). A smaller percentage of changes were the addition of new lines (13%) and a very low percentage were removals of lines (1%). The updates refer to the values to more than 85% and included the following operations: (i) deleting a value and adding a different one afterwards (because of disputes between authors) (ii) replacing or updating values (iii) changing the order of keys. An article about a controversial topic can cause large and intensive discussions with more than 240,000 words[2] and up to 40,000 revisions per article (e.g. the article about George W. Bush). On average, an article consists of 40-50 revisions.

Every registered Wikipedia user can easily modify the values of an infobox. In contrast, changing the keys and, thus, a template's structure is not easy. Many templates are locked by a Wikipedia administrator or have very long waiting queue for modifications because every user wants to add his personal ideas to the template.

The second analysis was focused on the templates them-

selves. The source codes of all templates were extracted from an English dump and the contained keys were compared. This analysis revealed that many templates use the same keys and that some of them match up to 100%, e.g. "Venus_crater" and "Mercury_Crater". They are redundant, can lead to confusions and should be merged. Moreover, Figure 1 shows that many templates are used very rarely, resulting in a longtail distribution. Some keys like "name", "image" or "website" are part of more than 60% of all templates. Another problem are synonym keys like "homepage" and "website". Such multiple usages of keys and their difficult modification makes the concept of templates highly inflexible.

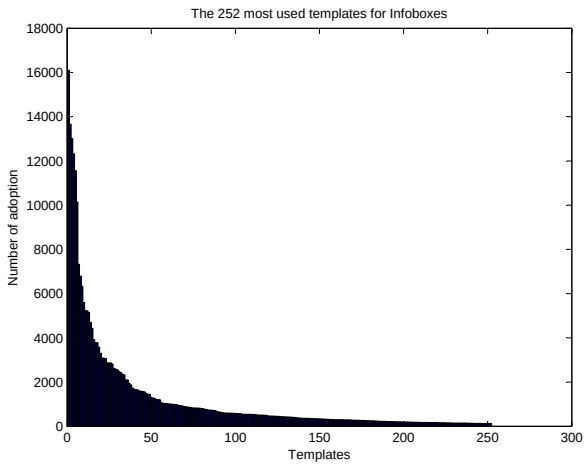

Figure 1: Most popular templates.

3. KEY RECOMMENDATION

Our new concept avoids the use of templates and introduces a self-learning system which does not need any administrators. SnoopBox controls itself by managing the keys entered by the user. Furthermore, it creates a ranking of these keys. Frequently used keys are ranked higher and recommended more often. The user is guided by the recommendations but keeps the freedom to choose his desired keys.

SnoopBox works with eight predefined categories and subcategories selected by the user. As a first step, the user has to assign a category to his new article. After this assignment, the system displays a list of already existing subcategories. The user can choose one of them or create a new one. Based on this selection, the system recommends the first keys. It displays the most popular keys from this subcategory, the most popular keys from this category and the overall most used keys. The user is now able to choose from this list and insert them into his new infobox. SnoopBox analyses the selected keys and "snoops" already existing infoboxes which contain a similar combination of keys. It takes keys from these infoboxes and recommends them to the current user. This process can be repeated until the final saving action. Throughout this process, a spell-checker avoids typos and resulting synonyms.

This recommendation system significantly simplifies the creation of infoboxes and reduces the amount of keys by avoiding the use of synonyms. Another benefit of this approach is the fact that the infoboxes are stored in the form of subject-key-value triples and therefore are available without any preperformed extraction process from the plain text. This approach makes it possible to query the structured data sets by using query-languages like, e.g. SQL or SPARQL to find appropriate keys and compare their values. This allows even complex queries like "Return all cities with a population higher than 100,000 which have a female mayor who has a doctoral degree". Furthermore, due to the more homogeneous data set, the quality of the query results is also increased.

Wikipedia		SnoopBox	
1. image	6. location	1. image	6. country
2. name	7. country	2. website	7. elevation
3. caption	8. imagesize	3. logo	8. capital
4. website	9. alt	4. location	9. inhabitants
5. type	10. logo	5. founded	10. type

Table 1: The top ten most used keys.

4. PRELIMINARY RESULTS

The evaluation of SnoopBox confirmed its functionality. Ten test-users got the task to create one or two articles containing an infobox and to fill-out a questionnaire afterwards giving their opinions about the tool. The majority of the test subjects approved the recommended keys and the tool itself. More than 90% would use it for infobox creation if it were available. The evaluation of the inserted keys showed that the top ten used keys covered the top ten keys in Wikipedia by about 60% as visible in Table 1. The newly created infoboxes covered comparable infoboxes in Wikipedia by more than 50%. These numbers are surprisingly high because most of the test-users were not familiar with the creation of infoboxes before this test. After the test, everyone was able to create an infobox. Furthermore, the users were confronted with the actual implementation in Wikipedia, but only 30% of them succeeded in creating a valid infobox using a template. These 30% corresponds to the percentage of participating computer scientists in the test group.

5. CONCLUSION

The analysis of the actual process of infobox creation and storage in Wikipedia revealed many flaws. The concept of templates is not very flexible and not easy to understand. Furthermore, the support provided to the user is minimal. SnoopBox is a self-learning user support tool using key recommendation and a more structured storage of data inside infoboxes. The evaluation showed that the users accept SnoopBox and would use it because they feel more confident when creating an infobox. The inserted keys and groups of keys covered those used in the original Wikipedia templates to a high grade after the insertion of only 20 articles. A potential application in Wikipedia is intended in the future.

6. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web, Lecture Notes in Computer Science*, 4825:722–735, November 2007.

[2] L. Gomes. Forget the articles, best wikipedia read is its discussions. *Wall Street Journal*, page B1, August 2007.

[3] F. Wu and D. S. Weld. Automatically refining the wikipedia infobox ontology. *Proceeding of the 17th international conference on World Wide Web*, 2008.