EachWiki: Suggest to Be an Easy-To-Edit Wiki Interface for Everyone

Huajie Zhang, Linyun Fu, Haofen Wang, Haiping Zhu,
Yang Wang, and Yong Yu

Dept. of Computer Science and Engineering, Shanghai Jiao Tong University
800 Dongchuan Rd, Shanghai, P.R.China, 200240
{zhjay, fulinyun, whfcarter, zhu, wwwy, yyu}@apex.sjtu.edu.cn

Abstract. In this paper, we present EachWiki, an extension of Semantic MediaWiki characterized by an intelligent suggestion mechanism. It aims to facilitate the wiki authoring by recommending the following elements: *links*, *categories*, and *properties*. We exploit the semantics of Wikipedia data and leverage the collective wisdom of web users to provide high quality annotation suggestions. The proposed mechanism not only improves the usability of Semantic MediaWiki but also speeds up its converging use of terminology. The suggestions are applied to relieve the burden of wiki authoring and attract more inexperienced contributors, thus making Semantic MediaWiki even better Semantic Web proto types and data source.

1 Introduction

As popular tools for collaborative authoring, wikis have been considered as effective content or knowledge management solutions. One of the best-known wikis is *Wikipedia*[1], the largest free online encyclopedia served by the famous wiki engine MediaWiki[2]. Wikipedia embraces the power of collaborative editing to harness collective intelligence. Semantic Wikis aim to make knowledge accessible to machines (e.g. agents, services) beyond mere navigation, by combining wikis and the Semantic Web technologies. A featured example is *Semantic MediaWiki*[3] that adds explicit semantics (relations and attributes) to links between pages [4].

However, high technical barriers still exist for inexperienced users to use such wikis to create or formalize their knowledge, thus deterring many domain experts from being involved in. As stated in [3], the heavy burden of up-building and maintaining such an enormous knowledge base still rests on a very small group of people, which is incompatible with the supposed "wiki way". In spite of their flexibility and autonomy, current Wikipedia and Semantic Wikipedia might still cause trouble to contributors, especially those newcomers: while authoring a wiki page, he/she often

[1] http://en.wikipedia.org
[2] http://www.mediawiki.org
[3] http://ontoworld.org/wiki/Semantic_Mediawiki

feels at a loss due to lack of knowledge about the existing information (e.g. links[4], categories) accommodated in the system. The quality of articles will be affected by such confusions in the following cases: 1) the case of missing annotations (e.g. missing links [1] or missing categories), which will affect the knowledge access and reuse; 2) the case when users annotate arbitrarily and unreasonably, which does harm to the consistency of content. While simple centralized policies would highly restrict users needlessly, a mechanism of suggestions based on the dataset would relieve the burden of the "small group" and help terminology convergence without forcing users [5]. To provide such a mechanism, we present EachWiki[5], an extension of Semantic MediaWiki that integrates the following suggestion modules into the editing interface: link suggestion, category suggestion, and property[6]suggestion. To summarize, the proposed suggestions are applied to 1) improve the usability of (semantic) wikis; 2) improve the quality of articles in the process of wiki authoring; 3) facilitate the creation of semantic descriptions about resources.

In the rest of this paper, we first enumerate the features of EachWiki in Section 2, then we describe the three suggestion modules in Section 3 to 5 accordingly, and finally, we give our conclusion and future work in Section 6.

2 Characteristics of EachWiki

The suggestions of EachWiki are featured by the following aspects:

AJAX-based and highly interactive user interface. EachWiki aims to attract more inexperienced domain experts to be involved in the process of knowledge creation. To provide real time suggestions without disturbing users unnecessarily, EachWiki resorts to AJAX technology to seamlessly integrate the search function into the editing interface.

Data manipulation – scalable to real web environments. We have crawled, extracted and indexed the entire English version of Wikipedia (by June, 2007), that is about 1,800,000 articles and 200,000 categories, to support the powerful suggestions, which can reflect the holistic view of Wikipedia.

Real time suggestions offering satisfactory efficiency. The suggestions have short response time, in spite of the large scale of the indexed data.

Full exploitation of the semantic features of Wikipedia data. We make full use of the semantic features of Wikipedia data and exploit the RDF graph matching techniques to perform filtering and ranking of results, which guarantees the high quality of suggestions.

Support of dynamic and incremental knowledge content creation. Knowledge evolves dynamically in Wikipedia: the versions of articles increase quickly. Hence, EachWiki aims to facilitate the revising and refining on existing pages: as the user

[4] We will simply refer to "internal links" as "links" in the rest of this paper, for detailed explanations see: http://en.wikipedia.org/wiki/Help:Interwiki_linking

[5] http://eachwiki.apexlab.org

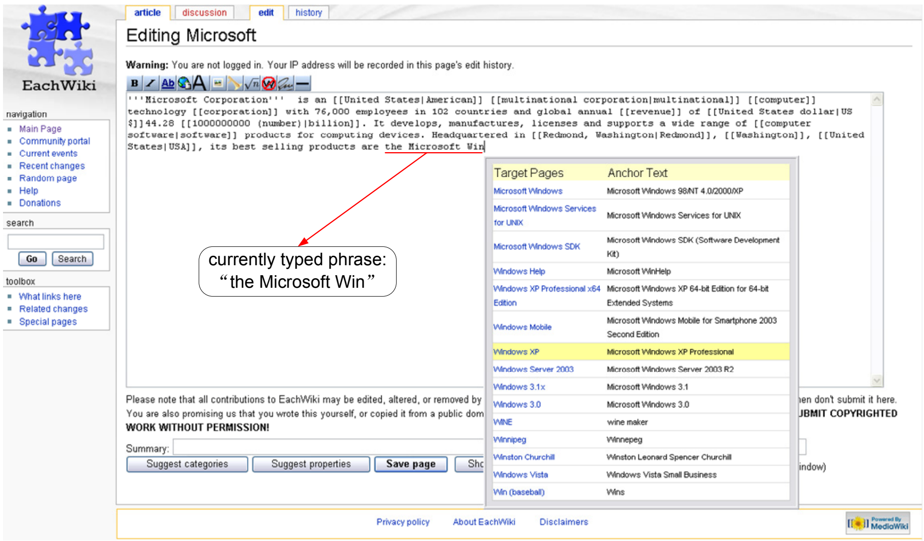
[6] In OWL, a property is a binary relation, including two types: *datatype property* (or *attribute*) and *object property* (or *relation*). In EachWiki, properties mean either of them, which depict explicit relations or attributes on links of wiki pages.

accepts the suggestions and modifies the content, the results of the new-round suggestions will be changed and refined accordingly.

For further understanding of the above features of EachWiki, its main building blocks – three suggestion modules – will be introduced in the following sections.

3 Link Suggestion

Links between pages allow the user to access information related to perform hyper-reading [8], which are created by surrounding words with double square brackets, and any spaces between them are left intact, e.g. "[[World Wide Web]]". Simple as the link syntax is, the situations of missing links in Wikipedia [1] usually happen when the author does not realize there should be a link. Sometimes even when the author is aware that a link is necessary, he/she might still fail to locate the corresponding target article when he/she uses an expression different from the title of that article (titles serve as IDs for retrieving articles). In such a case, the author tends to provide a red link which points to a page that does not actually exist in the system. A thoughtful user might take advantage of the search function provided in the system. Obviously, this is inconvenient and will disturb the authoring process.



Editing Microsoft

Warning: You are not logged in. Your IP address will be recorded in this page's edit history.

```
'''Microsoft Corporation'''  is an [[United States|American]] [[multinational corporation|multinational]] [[computer]]
technology [[corporation]] with 76,000 employees in 102 countries and global annual [[revenue]] of [[United States dollar|US
$]]44.28 [[1000000000 (number)|billion]]. It develops, manufactures, licenses and supports a wide range of [[computer
software|software]] products for computing devices. Headquartered in [[Redmond, Washington|Redmond]], [[Washington]], [[United
States|USA]], its best selling products are the Microsoft Win
```

currently typed phrase: "the Microsoft Win"

Target Pages	Anchor Text
Microsoft Windows	Microsoft Windows 98/NT 4.0/2000/XP
Microsoft Windows Services for UNIX	Microsoft Windows Services for UNIX
Microsoft Windows SDK	Microsoft Windows SDK (Software Development Kit)
Windows Help	Microsoft WinHelp
Windows XP Professional x64 Edition	Microsoft Windows XP 64-bit Edition for 64-bit Extended Systems
Windows Mobile	Microsoft Windows Mobile for Smartphone 2003 Second Edition
Windows XP	Microsoft Windows XP Professional
Windows Server 2003	Microsoft Windows Server 2003 R2
Windows 3.1x	Microsoft Windows 3.1
Windows 3.0	Microsoft Windows 3.0
WINE	wine maker
Winnipeg	Winnepeg
Winston Churchill	Winston Leonard Spencer Churchill
Windows Vista	Windows Vista Small Business
Win (baseball)	Wins

Please note that all contributions to EachWiki may be edited, altered, or removed by … hen don't submit it here.
You are also promising us that you wrote this yourself, or copied it from a public dom … UBMIT COPYRIGHTED
WORK WITHOUT PERMISSION!

Summary:

Figure 1. Editing interface of EachWiki with link suggestion view in it.

To avoid disturbing users unnecessarily in the process of authoring, the link suggestion should be performed in real time. Therefore, EachWiki resorts to AJAX technology to seamlessly integrate the search function into the editing interface. As the user types a phrase (can be incomplete) in the editor and pauses for a short time, the link suggestion module will be triggered and the suggested results will be popped up. Figure 1 shows the editing interface of link suggestion view: the table below the currently typed phrase "the Microsoft Win" contains a list of suggested links corresponding to it. This module aims to "guess" what resources the user is typing and suggest them in real time. In the list, each row represents a link having two columns as: the first column being the title (also the unique ID) of the target article that is linked to (e.g. "Windows XP"), whereas the second column being the most frequently annotated anchors of that link (e.g. "Microsoft Windows XP Professional"). After the user selects and confirms anyone in the list by pressing the return key or double-clicking the mouse, the corresponding link annotation is generated to replace the original typed phrase automatically. For example, after the user selects the link "Windows XP", the typed phrase "Microsoft Win" will be replaced by link annotation with its anchor as: "[[Windows XP| Microsoft Windows XP Professional]]".

In link suggestion, the search is similar to the auto-completion search in *Google Suggest*[7], but more sophisticated. Firstly, the query phrase is not the entire string in the text area. Each time the link search module is activated, the system gets the current phrase by looking backwards k (currently we make k=3) words from the current cursor position and try its different suffix sub-phrases as query phrases for prefix search. In the above example, we not only get results matched by "Win*" (e.g. "Windows Vista"), but also get ones matched by "Microsoft Win*" (e.g. "Microsoft Windows"). Secondly, the search for each query is more than just prefix matching over page titles. Since title is the main building block of Wikipedia vocabulary, most, if not all, of the existing Wikipedia search engines or authoring assistants, such as *LuMriX*[8], *WikiWa*x[9] and *Plog4U*[10], support only prefix matching over article titles. However, it is imaginable that the anchor text of a piped link[11] can be totally different from the title of the target article. In EachWiki, matching and ranking of links are supported by the following evidences for full exploitation of the meaning of Wikipedia thesaurus:

- **Title**: The title serves not only as the name of the encyclopedia entry but also as the unique resource ID for it.
- **Redirect**: All the titles redirecting to a given article are synonyms of it, e.g. "The Web" and "WWW" redirects to article "World Wide Web".
- **Disambiguation**: A disambiguation page is a page containing paths (links) leading to the different article pages that could use essentially the same term as their title, which is a polysemy term, e.g. in page "WWW (disambiguation)", the term "WWW" is a polysemy term that might refer to article "World Wide Web" or film "Wild Wild West".
- **Emphasized Phrase**: Generally in each article of Wikipedia, the first sentence serves as the definition of the *principal entity*[12]. The emphasized bold phrase in the definition sentence is a self-reference to the principal entity. It is considered as the synonym of principal entity or the full name (or abbreviation) of a person.
- **Anchor Text**: Anchor text is the displayed text of a link in pages. For example, while in simple link "[[World Wide Web]]", the anchor text for "World Wide

[7] http://www.google.com

[8] http://wiki.lumrix.net/en

[9] http://www.wikiwax.com

[10] http://www.plog4u.org/index.php/Main_Page

[11] http://en.wikipedia.org/wiki/Piped_Link

[12] In encyclopedia such as Wikipedia, the articles are biographical article. A biographical article mostly focuses on only one entity that is reffered as the "principal entity".

Web" its title "World Wide Web", in piped link "[[World Wide Web|WWW]]" , the anchor text for it is "WWW".

The above semantic features of Wikipedia articles enable the link suggestion module to find relevant articles with titles possibly different from the typed phrases. For example, when the user types "William Henry Gates III", the article "Bill Gates" can be returned (for it is the full name of Bill Gates), which is beyond the capability of a simple title prefix match.

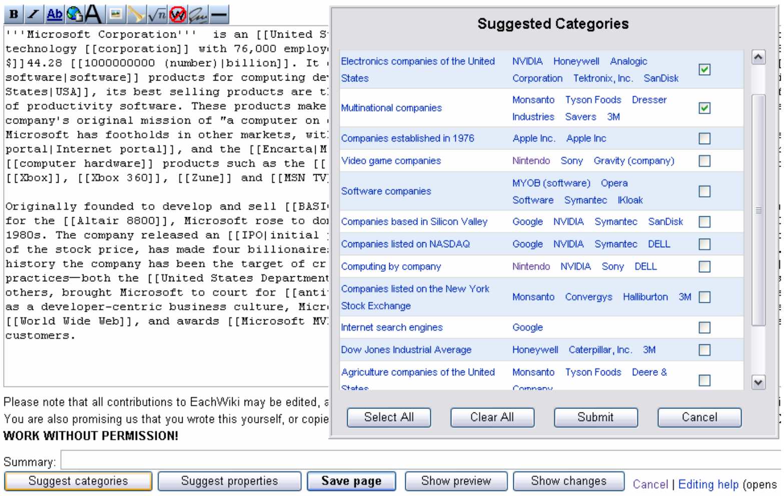
4 Category Suggestion

A category is a collection of articles classified according to their content. Categories are organized in a hierarchical structure so that users can navigate from a category to its sub-categories or super-categories. In such a fashion, users manage to narrow (or expand) their browsing scope to more specialized (or more generalized) topics. In a word, categories are important elements because: 1) they are summarization of articles; 2) they facilitate the knowledge accessing and browsing. Hence it has become the major concern how to make high-quality category annotations to put pages where they should be. When categorizing a newly created article, due to lack of the whole picture of the category structure, the user (esp. a novice) might feel at a loss whether to reuse an existing category or create a new one.

EachWiki can reduce such confusion by recommending relevant categories. On the other hand, even when the user annotates some proper categories, it is usually difficult for the individual to annotate a set of categories with a high coverage of multi facets of the resource. For example, while many people know that Bill Gates is one of famous "American entrepreneurs" and "American billionaires", not everyone knows that he is also one of "Dropouts of Harvard University" and "American agnostics", which describe other facets (categories) of Bill Gates. In this case, category suggestion is also of great value during the revising of an existing article, because it gives guidance to append missing categories, which help users refine the existing pages incrementally and improve the coverage of its categories.

Similar to the thought of *collaborative filtering* [6] that uses the ratings from other like-minded users to calculate a prediction for the active user (the user whom the prediction is for), the mechanism of category suggestion is to predicate the relevant categories of the target article based on the ones from similar articles, which takes two steps: 1) discover similar articles and 2) rank their categories. Our observation is that articles sharing some sets of features usually tend to share other sets of features. For example, articles in the same category usually share some features such as infoboxes, section headings, and they often link to and are linked by some common articles. The similarity between any two articles is defined by the extent to which they share their features such as: the articles linking to the current article, articles linked by the current article, the section headings of the article (indicating the organization structure of the page), categories of the article, etc. Detailed algorithm is described in [7].

Once pressing the first button "Suggest categories" on the bottom, the user activates the action of category suggestion, the results of which are shown in the table (in Figure 2). The table in the figure shows a ranked list of categories that are most related to the created article, along with the corresponding evidences (from which relevant articles the suggested categories come, e.g. the similar articles of Microsoft are: “NVIDIA”, “Honeywell”, “Analogic”, etc.). The user can accept any of them by selecting the checkboxes on the right side of the corresponding rows. After submission, the wiki code of category annotation (e.g. “[[Category:Multinational Companies]]”) is generated and appended into the text area



```
'''Microsoft Corporation'''  is an [[United S
technology [[corporation]] with 76,000 employ
$]]44.28 [[1000000000 (number)|billion]]. It
software|software]] products for computing de
States|USA]], its best selling products are t
of productivity software. These products make
company's original mission of "a computer on
Microsoft has footholds in other markets, wit
portal|Internet portal]], and the [[Encarta|M
[[computer hardware]] products such as the [[
[[Xbox]], [[Xbox 360]], [[Zune]] and [[MSN TV

Originally founded to develop and sell [[BASI
for the [[Altair 8800]], Microsoft rose to do
1980s. The company released an [[IPO|initial
of the stock price, has made four billionaire
history the company has been the target of cr
practices—both the [[United States Department
others, brought Microsoft to court for [[anti
as a developer-centric business culture, Micr
[[World Wide Web]], and awards [[Microsoft MV
customers.
```

Please note that all contributions to EachWiki may be edited, a
You are also promising us that you wrote this yourself, or copie
WORK WITHOUT PERMISSION!

Summary:

Suggested Categories

Electronics companies of the United States	NVIDIA Honeywell Analogic Corporation Tektronix, Inc. SanDisk	☑
Multinational companies	Monsanto Tyson Foods Dresser Industries Savers 3M	☑
Companies established in 1976	Apple Inc. Apple Inc	☐
Video game companies	Nintendo Sony Gravity (company)	☐
Software companies	MYOB (software) Opera Software Symantec IKloak	☐
Companies based in Silicon Valley	Google NVIDIA Symantec SanDisk	☐
Companies listed on NASDAQ	Google NVIDIA Symantec DELL	☐
Computing by company	Nintendo NVIDIA Sony DELL	☐
Companies listed on the New York Stock Exchange	Monsanto Convergys Halliburton 3M	☐
Internet search engines	Google	☐
Dow Jones Industrial Average	Honeywell Caterpillar, Inc. 3M	☐
Agriculture companies of the United States	Monsanto Tyson Foods Deere & Company	☐

Figure 2. Editing interface of EachWiki with category suggestion view in it.

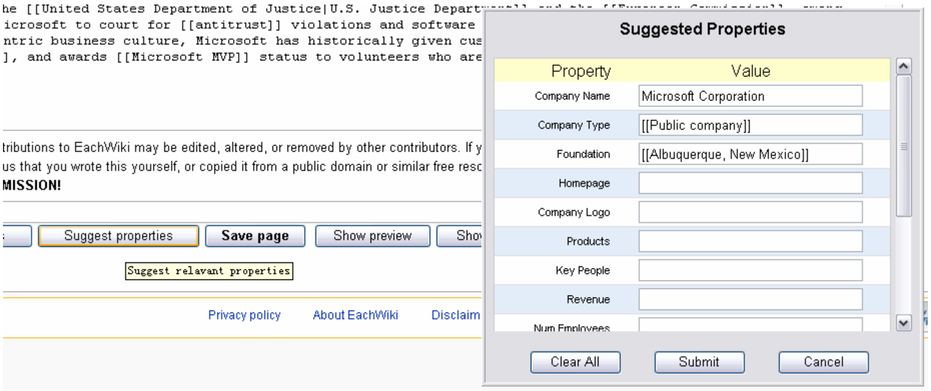
5 Property Suggestion

In addition to links and categories, Semantic Wikis also enable users to make semantic statement (RDF triple) arbitrarily. However, the authoring freedom in Semantic Wikis would result in statements with different vocabularies, violating the purpose of the Semantic Wiki. Therefore, EachWiki provides property suggestion based on the structured data currently available in Wikipedia: the tabular information contained in Wikipedia *Infoboxes*[13]. An infobox describing a class of entities contains a set of items, each of which is represented in a property-value notation and can correspond to an RDF triple statement with the principle entity as its subject, the property as its predicate, and the value (entity or data type) as its object. For example, article “Bill Gates” has an infobox “Infobox Person” that contains some properties of “Person” such as “Born”, “Occupation”, etc. The “Occupation” property with a hyperlink form “[[Chairperson|Chairman]]” as its value describes an “Occupation” relation associating “Bill Gates” with “Chairperson”. Another property “Net Worth” with plain text

[13] Wikipedia enables authors to include predefined content or display content in a determined way, which is realized by the Template. As a special case, Infobox aims to generate consistently-formatted boxes in a tabular form.

value "US$56 billion" can be regarded as an attribute of Bill Gates. As supposed in [2], approximately a quarter to one third of the Wikipedia pages today already contain such kind of structured information, which is valuable for querying and machine interpretation and qualifies as an important data source for property suggestion. In addition to infobox data, the relation and attribute annotations on links can also be considered as another data source for property suggestion in the future, although they are currently not accommodated in Wikipedia and have not been indexed yet.

This module works in the same way as the category suggestion module: it uses similar features, and it adopts similar searching and ranking techniques. As in Figure 3, after the user presses the button "Suggest properties", a list of suggested properties emerges. The value (or object) of each property is left empty for the user to fill in. Once the user keys in any of them and submits, the corresponding wiki text of the relation or attribute annotations will be generated and inserted into the text area (the syntax is defined in Semantic Wikipedia [4]): while relation annotations are given on the links in the input box, attribute annotations are given on plain literals, each of which corresponds to an RDF triple and can be explored, exported and retrieved.



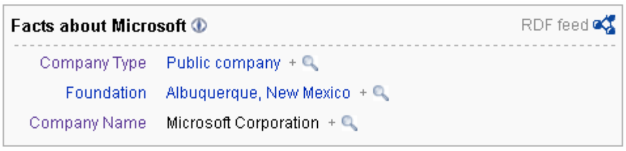
```
he [[United States Department of Justice|U.S. Justice Depart
icrosoft to court for [[antitrust]] violations and software
ntric business culture, Microsoft has historically given cus
], and awards [[Microsoft MVP]] status to volunteers who are
```

tributions to EachWiki may be edited, altered, or removed by other contributors. If y
us that you wrote this yourself, or copied it from a public domain or similar free reso
MISSION!

Suggested Properties

Property	Value
Company Name	Microsoft Corporation
Company Type	[[Public company]]
Foundation	[[Albuquerque, New Mexico]]
Homepage	
Company Logo	
Products	
Key People	
Revenue	
Num Employees	

Figure 3. Editing interface of EachWiki with property suggestion view in it.



Facts about Microsoft

Company Type	Public company +
Foundation	Albuquerque, New Mexico +
Company Name	Microsoft Corporation +

Figure 4. Content page view of the annotated properties.

For example, after the user creates (or revises) the page "Microsoft" and presses "Suggest properties" button, the returned properties are as follows: "Company Name", "Company Type", "Foundation", etc. The user can choose any of them and finish the triples by adding wiki code in their object position. For example in Figure 3, the user has input a literal String "Microsoft Corporation" in the "Company Name" field, which is then stored as attribute annotation like "[[Company Name:= Microsoft Corporation]]"; and he/she has also typed a link annotation "[[Albuquerque, New Mexico]]" in the "Foundation" field, which is recognized as a relation annotation "[[Foundation::Albuquerque, New Mexico]]". After the system automatically converts the property-value items into relation or attribute annotations, the triples in the content page view are shown in Figure 4.

In a word, property suggestion module aims to facilitate the creation of arbitrary RDF statements of resources, which are considered as consistent Semantic Web data. It helps the semantic reuse and promotes the terminology convergence in Semantic Wikis.

6 Conclusion and Future Work

In this paper, to facilitate the knowledge creation, access and reuse in wiki environments, we present EachWiki, an extension of Semantic MediaWiki equipped with a suggestion mechanism to suggest links, categories and properties. It aims not only to relieve the burden of Wikipedia authoring but also to attract a broader community by making more (Semantic) Wikipedians.

Our future work includes: First, large-scale user evaluations will be carried out to prove the effectiveness and efficiency of our suggestion modules. Second, the usability will be improved by providing proper guidelines for users to fill in the values of properties in the property suggestion module. Third, we plan to exploit more semantics in (Semantic) Wikipedia for other suggestion modules.

References

[1] Adafre, S.F., Rijke, M.: Discovering Missing Links in Wikipedia. LinkKDD 2005.

[2] Auer, S., Lehmann, J.: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. ESWC 2007.

[3] Swartz, A.: Raw Thought: Who Writes Wikipedia. http://www.aaronsw.com/weblog/whowriteswikipedia

[4] Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. WWW 2006.

[5] Oren, E., Gerke, S., Decker, S.: Simple Algorithms for Predicate Suggestions using Similarity and Co-Occurrence. ESWC 2007.

[6] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews. CSCW 1994.

[7] Wang, Y., Wang, H., Zhu, H., Yu, Y.: Exploit Semantic Information for Category Annotation Recommendation in Wikipedia. NLDB 2007.

[8] Zhang, Y.: Wiki Means More: Hyperreading in Wikipedia. HT 2006.