EachWiki: Facilitating Semantics Reuse for Wikipedia Authoring
Haofen Wang, Huajie Zhang, Linyun Fu and Yong Yu
Dept. of Computer Science and Engineering, Shanghai Jiao Tong University
800 Dongchuan Rd, Shanghai, P.R.China, 200240
{whfcarter, zhjay, fulinyun, yyu}@apex.sjtu.edu.cn

ABSTRACT
Wikipedia has grown into the largest free online encyclopedia as it follows the “wiki way”. The promising roadmap of Wikipedia has also aroused the great interests from the Semantic Web community. However, the heavy burden of upbuilding and maintaining such an enormous knowledge base still rests on a very small group of people. In this paper, we propose an integrated solution to facilitate semantics reuse for Wikipedia authoring. Such semantics reuse comes from providing users with: 1) A link suggestion module that seamlessly integrates search and authoring to reuse existing article hyperlinks; 2) A category suggestion module that borrows relevant categories from similar articles; 3) A semantic relation suggestion module that facilitates annotating candidate triples by reusing existing triple types. The whole approach can be applied to enhance the (Semantic) Wikipedia. A prototype system named EachWiki is implemented and presented together with its editing interface.

Author Keywords
Wikipedia, Semantic Wikipedia, link suggestion, category suggestion, semantic relation suggestion

ACM Classification Keywords
H.5.2. Information interfaces and presentation: Evaluation/methodology, Interaction styles, Prototyping, Screen design, User-centered design.

INTRODUCTION
As popular tools for collaborative authoring, wikis have been considered as effective knowledge management solutions, which prefer community-centered efforts to private, personal web sites. One of the best-known wikis is Wikipedia, the largest free online encyclopedia authored by a broad community of volunteers. Wikipedia embraces the power of collaborative editing to harness collective intelligence. The promising roadmap of Wikipedia aroused the great interests from the Semantic Web community. Actually, if we treat categories as classes, articles as class instances, it is natural to look Wikipedia as an inherent taxonomic system [11]. To make a step further, an intuition is to add explicit semantics (relations or attributes) to links of Wikipedia. The Semantic Wikipedia [6] gives a comprehensive blueprint of such enhancement by combining wikis and the Semantic Web technologies. It aims to make knowledge accessible to machines (e.g. agents, services) beyond mere navigation.

Recent research [4] has suggested that the accuracy of the Wikipedia is not much worse than that of the editorially complied encyclopedias such as Encyclopedia Britannica. The surprisingly high quality and popularity of Wikipedia should be owed to the so-called “wiki way” in constructing the possibly largest knowledge base all over the world. We may characterize it simply as openness, collaboration, democracy and free of hard restrictions. However, an interesting (and somewhat worrying) finding has been reported that 80% of the articles are edited by only 10% contributors [12]. Wales, the founder of Wikipedia, even pessimistically declared that the most active 2%, which is 1400 people, did 73.4% of all the edits, while the remaining 25% of edits were from people who are contributing a minor change of a fact, a minor spelling fix or something like that [4]. We may reckon that the Semantic Wikipedia is likely to face the similar problem. That is to say, the heavy burden of upbuilding and maintaining such an enormous knowledge base still rests on a very small group of people, which is incompatible with the supposed “wiki way”.

It can be observed that while authoring a Wikipedia article, the user might usually feel at a loss due to lack of knowledge about the existing semantic information accommodated in the system. For instance, when is it necessary to provide a hyperlink to a related topic article for reference? What categories are proper to characterize an article? And especially for a Semantic Wikipedia article, -- what semantic relations are required to model the properties and the attributes of an entity? A real anxiety is that the user must be very cautious about these questions in order to assure the quality of the article that he/she is authoring. Unfortunately, however, collaboration seems to only come into favor with the next edit of the same page.
Consequently, if we integrate the reuse of the above-mentioned semantics into the Wikipedia authoring interface, the “wisdom of crowds” will be brought to the front. When the collaboration of quality assurance is concretized within each edit of an article, more casual users are expected to produce high-quality outputs as easily as advanced users, and thus be willing to dedicate themselves to Wikipedia.

To summarize, we propose in this paper a novel solution to equip the current (Semantic) Wikipedia with:

- **A link suggestion** module that seamlessly integrates search and authoring to reuse existing article hyperlinks;
- **A category suggestion** module that borrows relevant categories from similar articles;
- **A semantic relation suggestion** module that facilitates annotating candidate triples by reusing existing triple types.

The whole approach can be applied to enhance the current Wikipedia and the Semantic Wikipedia, as well as to bridge from the former to the latter. We expect that the implementation is able to be introduced into Wikipedia in the near future. We implement a prototype system called EachWiki \(^1\) which gives a testing workbench of the proposed enhancements.

The remainder of the paper is organized as follows. Section 2 elaborates our solution of facilitating semantics reuse over links, categories and semantic relations in such free encyclopedic environments. Section 3 discusses some related work. Section 4 concludes the whole paper.

**OUR SOLUTION**

As mentioned in Section 1, **links, categories and semantic relations** are among the typical examples that the users would like to seek for system support while authoring a (semantic) Wikipedia article. Also, we are inspired by the “k-NN” thought that the usage pattern in similar contexts would suggest the use of an element in a repeatedly appeared context. Our solution has several attractive characteristics to fulfill the above suggestion tasks:

- **Effectiveness.** Different kinds of semantic features in the (Semantic) Wikipedia are exploited to guarantee not only the high quality but also a good coverage of suggestion results. 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;
- **Efficiency.** The famous IR engine Lucene \(^2\) is leveraged to support indexing and searching the underlying large scale (Semantic) Wikipedia data, which guarantee short response time;
- **Usability.** AJAX technology is seamlessly integrated into the editing interface to provide highly interactive user interface of suggestions without disturbing users unnecessarily.

In this section, a unified suggestion model will be proposed first, the suggestion mechanisms on the three elements are then to be explained in detail according to the unified suggestion model.

**Unified Suggestion Model**

Inspired by the thought of collaborative filtering \(^8\) 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 classification-based suggestion is to predicate the relevant features for the target resource based on the ones from similar resources, which takes two steps: 1) a classification step to discover similar resources and 2) a ranking step to rank their features. The algorithm is described in Figure 1:

![Algorithm: suggest (query \(q\), resource set \(E\)](attachment:Algorithm.png)

**Link Suggestion**

The situations of missing links \(^1\) 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 page when he/she uses an expression different from the title of that page (titles serve as IDs for retrieving the resources). 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 to find fuzzy matching of the expression in any **title**, **body text** or **category** field of an article. Obviously, this is inconvenient and will disturb the authoring process. Thereupon, **what if we seamlessly integrate the
search function into the editing interface? The success of auto-completion products has given the positive evidence.

We propose an information retrieval approach to link suggestion. The search is similar to the auto-completion search in Google Suggest3, but more sophisticated. Firstly, the query phrase is not the entire string in the text area. Each time the link suggestion 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. When typing “Win”, 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 LuMriX4, WikiWax5 and Plog4U6, support only prefix matching over article titles. However, it is imaginable that the anchor text of a piped link can be totally different from the title of the target article. Each article is represented by a virtual document containing the following fields, each of which is a bag-of-words vector:

- **Title**: the unique ID for the given Wikipedia article, e.g. “World_wide_web” for “World Wide Web”.
- **RedirectTo**: synonym phrase that shares the sense with the title, e.g. “The_Web” redirects to “World_Wide_Web”.
- **Disambiguation**: document title from different disambiguation pages, e.g. “WWW (disambiguation)” can refer to “World_Wide_Web”.
- **Emphasized Phrase**: The bold emphasized phrase in the first sentence of the article. e.g. “WWW” for “World Wide Web”.
- **Anchor Text**: Each page may, however, be represented by different anchor texts in different hyperlinks. For example, “Web” and “WWW” both link to “World Wide Web” but with different anchors.

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.

The corresponding similarities function in the classification step is defined as follows:

\[
sim(q,e) = \sum_{field \in d} weight(field) \times score(q, field)
\]

\[
score(q, field) = \sum_{term \in field} pmatch(q,term) \times tf(term) \times idf(term) \times LengthNorm(field)
\]

\[
mpmatch(x, y) = \begin{cases} 1, & \text{if } x \text{ is prefix of } y; \\ 0, & \text{otherwise}. \end{cases}
\]

One important design principle of link suggestion is that we should give higher scores to links with popular anchors, so we set the length normalization function \(LengthNorm(field) = 1\), which means the more times the query prefix match the terms, the score will be higher.

In the ranking step, the candidates are simply the titles (ID) of the relevant resources: \(Candidate = \{e.title \mid e \in R(q)\}\). The ranking function is: \(RankFunc(e.title) = sim(q,e)\), which means that the titles are ranked according to the similarity between the corresponding articles and the query.

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 2 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]]”.

![Figure 2. Editing interface of EachWiki with link suggestion view.](image)

---

3 http://www.google.com
4 http://wiki.lumrix.net/en
5 http://www.wikiwax.com
6 http://www.plog4u.org/index.php/Main_Page
Category Suggestion

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. Category suggestion can reduce such confusion by recommending relevant categories for the article. 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 multiple 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.

Our observation is that articles sharing the same category of features usually tend to share other sets of features. For example, articles in the same category usually share some features such as infobox names, section headings, and they often link to or are linked by some common articles. Then our concern is how to define the similarity function and the ranking function in the suggestion model.

In the classification step, the query is an article (whether it is finished or not). We also use the virtual document to model the query article or resource article. In addition to categories, other important features considered are as follows:

- **Section Heading**: Sections structurally denote sub-topics which depict some attributes of current article. Especially, the section headings are the most representative identifiers of these sub-topics.

- **Infobox Name and Infobox Property**: In Wikipedia, there is the tabular information contained in Infoboxes7. An infobox describing a class of a certain resources contains a set of items, each of which is represented in a property-value notation. For example, in the content of article “Bill Gates”, there is an infobox named “Infobox Person” containing some properties of class “Person” such as “Born”, “Occupation”, etc. The infobox name “Person” denotes the class described, which means that “Bill Gates” is an instance of class “Person”. The “Occupation” property with a hyperlink form “[[Chairperson/Chairman]]” as its value describes a relation named “Occupation” and associating “Bill Gates” with the article “Chairperson”.

- **InLink and OutLink**: While InLinks are articles linking to the current article, OutLinks are articles that current article links to. Inter-links in Wikipedia articles build up an implicit semantic network, thus by clicking the hyperlinks user is shifted to other articles with same or close topic. The phenomenon indicates inter-links imply the semantic relations between two article instances.

Hence, the similarity function is defined as follows:

\[
\text{sim}(q,e) = \frac{\sum \text{weight}(field) \times \text{score}(d_q, field)}{\sqrt{\sum \text{score}(d_q, field)^2}}
\]

\[
\text{score}(d_q, field) = \sum \text{tf}(t) \times \text{idf}(t)^2 \times \text{lengthNorm}(field)
\]

The method is adaptive to different domains. It also has a few valuable characteristics:

**The method is able to suggest missing categories**. For example, in the suggestion list for “United Kingdom” there exists a category named “Category:Island nations”, whose evidence consists of “Republic of Ireland”, “Australia”, etc. The category does not exist in the page for “United Kingdom” in wikipediaXML, but is later appended to its updated version (with a synonymous category name “Category:Island countries”) in the current Wikipedia.

**The method is capable of discovering improper categorization**. For instance, in the experiment on “Web Ontology Language”, the category “Category:XML-based programming languages” was always “missing” (according to the ground truth), whatever the feature combinations were adopted. In fact, however, it is not necessary that Web Ontology Language (i.e. OWL) links to an XML-based markup language. The elimination of the category assignment in the current Wikipedia page for this concept appears to support our viewpoint as well.

**The algorithm can categorize an article to the proper level of abstraction**. Take the article for “Support Vector Machine” as an example, our algorithm suggests both “Category:Machine Learning” and “Category:Artificial Intelligence”, with the former ranked higher than the latter. By consulting the category hierarchy and the standard answer we found that “Category:Machine Learning” is subsumed by “Category:Artificial Intelligence” and the former should be associated with the article, which shows that our algorithm prefers to categorize an article to its immediate categories rather than their ancestors.

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 3). 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

---

7 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.
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.

Figure 3. Editing interface of EachWiki with category suggestion view.

Semantic Relation 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. We intend to go a step further to supply a reasonable reuse mechanism so as to normalize the semantic relations that are extracted from (the semantic) Wikipedia. We provide semantic relation suggestion based on the structured data currently available in Wikipedia: the tabular information contained in Wikipedia Infoboxes.

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]” as its value describes an “Occupation” relation associating “Bill Gates” with “Chairperson”. Another property “Net Worth” with plain text value “US$56 billion” can be regarded as an attribute of Bill Gates. As supposed in [3], 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 semantic relation 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 functions. As in Figure 4, 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 6): 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.

Figure 4. Editing interface of EachWiki with semantic relation suggestion view.

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 4, the user has input a literal String “Microsoft Corporation” as its object position. In a word, semantic relation suggestion module aims to facilitate the creation of arbitrary RDF statements of resources, which are considered as consistent Semantic Web

---

8 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.
data. It helps the semantic reuse and promotes the terminology convergence in Semantic Wikis.

RELATED WORK
Wikipedia has been greatly successful for its open collaborative principle. It has attracted much attention from the research community due to its abundance, influence, high quality and well-structuring. For example, [2] proved that the URIs of Wikipedia articles are surprisingly reliable identifiers for ontology elements; [13] extracted topics and their semantic relations from Wikipedia; [10] presented an experiment to automatically annotate several semantic relationships in Wikipedia; [9] presented measures for automatic filtering of strong semantic connections between Wikipedia categories; and [6] provided an extension to be integrated in Wikipedia that allows the typing of links between articles and the specification of typed data inside the articles in an easy-to-use manner.

Our work is mainly motivated by [4] and [5], which pointed out that Wikipedia’s authoring interface was not so convenient and smart as had been declared and that Wikipedia only had a relatively regular and small number of community members to take charge of most revision work. We hope to change such a situation. Research on Wikipedia Modeling is closely related to our work. [14] used several features of Wikipedia articles to extract concepts and recognize hierarchical relations between them.

Not much work has been done on the task of link suggestion in the Wikipedia environment. [1] addressed the problem of discovering missing links in Wikipedia. Their work concentrated on improving the existing link structure of Wikipedia and was not integrated into the Wikipedia authoring interface. Other related work includes several Wikipedia search engines, which only support prefix search over article titles, as mentioned in Section 2.2. A straightforward solution to the Wikipedia category suggestion task is to use the free text to represent Wikipedia articles as term weight vectors, measure similarity between two articles by the cosine of the angle between their vectors, and suggest categories.

In conclusion, we integrate link, category and semantic relation suggestions into authoring process, providing intelligent search functionality for Wikipedia and Semantic Wikipedia. The brand-new authoring process not only sweeps barriers of editing high-quality articles for causal wikipedians but also relieves heavy burden of Wikipedia administrators. We have designed and implemented an intelligent Wikipedia editor “EachWiki” to facilitate semantics reuse beneficial for both Wikipedia community and the Semantic Web community.

REFERENCES