# Analyses on Semantic Wiki and its Development

**Ji Zhao**
University of Auckland computer Science department
Zji002@ec.auckland.ac.nz

## ABSTRACT

Wiki is a kind of software that allows user to create and edit certain contents of website. It is often used to create collaboration on websites and to power community websites. Wiki lets public participant the website creation and maintenance. This feature decrease the cost of a website while increase the fame to the website audience, because people attend the Wiki website creation and maintenance reduce the cost and recommend that to users have same interests. It is so convenient and is expected more, but its nature limits the usage. It is just a collection of hypertext, and it neither supports structured access nor information reuse. The "data" contained in Wiki pages cannot be investigated by user or machine easily, unless they are sorted and stored in a special way. In order to overcome this, semantic technology is introduced to Wiki. A semantic Wiki basically is Wiki have underlying models, so the data and related data are managed in a good manner. Because of this feature, data can be investigated by user or machines easier and more efficiently. This seminar interests on semantic Wiki, as well as relative approaches, and try to find out future directions of semantic Wiki developing, as well as possible approaches.

## KEYWORDS

Wiki, semantic wiki, type link, attribute

## 1. INSTRUCTION

Wiki is a phenomenon on net now. The number of Wiki sites and its users increase a lot in recent years. At WikiSym 2005, Ward Cunningham and Jimmy Wales provided some answer: *"a wiki is like a garden; users (…) must take care of it. Start with some seeds and watch it grow, and the wiki will become moderated by its users' community, (…) respect and trust the users, (…) good things happen when you trust people more than you have reason to, let everybody express his opinion, no censorship, consensus must be reached, (…) the wiki is adapted toa dynamic social structure because of its refactoring features (…) Do not impose a rigid structure, users will refactor and structure the wiki as it grows (…)"* [1, 2, 3]. That is, wiki treat user as partners. Users' participant on creation and maintenance reduce the cost on creating and maintenance. Another advantage is increase the fame of the wiki site – the users will introduce the site to peoples in common as well. People normally recommend stuff they have confidence, and the attendance increase the confidence on the wiki site. Another feature is wiki is easy to be used. The creating and updating action is as easy as post and edit a message on net, and the content are formatted in a good manner. So peoples with little knowledge on net can share their knowledge on net, especially some experts of certain area.

Although wiki have many advantages, it is not convenient sometimes. The main problem complained most is the open structure makes navigation, orientation and search difficult, and wiki often fails to scale with the number of pages [4]. People expect more from wiki – the data should be able to be understood and investigated conveniently and effectively by both user and machine. In order to solve this, semantic technology is introduced to wiki.

Wikipedia define Semantic technology as "*technology encodes meanings separately from data and content files, and separately from application code*". Because wiki normally is a collection of hypertext, this technology will enrich Wiki content –the data contained in wiki is not only simple text, but can be used for further use. As a result, semantic wiki was developed.

Semantic Wiki is defined as "*A wiki that has an underlying model of knowledge described in its pages (…) .Semantic Wikis allow capturing or identifying further information about the pages (metadata) and their relations. Usually this knowledge model is available in a formal language, so that machines can (at least partially) process it*" in wikipedia. That is, the data are created, maintained and managed in certain models, which will simplify and enhance the search, navigation, as well as other processes. Compare to wiki, semantic wiki use a set of tables to store the information contained in plain text, so that queries and investigations an be performed more efficiently and easily.

In section 2 analyses will be held on semantic wiki, as well as its new features. In section 3 the possible approach to build a semantic wiki will be discusses, and section 4 to point out future directions of semantic wiki development in my opinion.

## 2. THEORETICAL ANALYSES

For the ease of discuss, assume there is a Wiki web site. This wiki is an encyclopaedia of movies and related information such as movie companies, directors and so on. Users can update and create movie information freely. And some of these processes need to be checked by an administration before they can be viewed from net. All sorts of information are identified by names.

## 2.1 Theroy approach

As a traditional wiki, it is easy for user to find out one movie if they know the name, and maybe it is easy to get a list of movie whose names contain certain word or letter. But it is hard to get a list of movie directed by a director, or published in a certain period or area – search need to be performed through the content of each piece of information. The cost is high—search through content increase the area need to be look up. The accuracy cannot be permitted as well, because the search can only tell the content contain the keyword, not the meaning of the keyword. For example, when you want to search movies directed by Chaplin, the search result will contain movies Chaplin stared. Adding abstract to each piece of information may reduce the work load, but it increases the work load of creating and updating as well.

In order to solve those, semantic wiki is developed. Basically it introduces models to manage the meanings in the information. Authors of [4] use wikipeadia as an example. In their design Meaning is divided to three parts – category, type link and attributes. Type link can be treated as a link with description [4], that is change a link from [[link article]] to [[type of link:: link article]]. For example, the sentence "… is directed by Charles Chaplin" will be implemented to "… is directed by [[Charles Chaplin]] in wiki, while in semantic wiki it will be "…is directed by [[is directed by::Charles Chaplin]] or even ignore the words before the "[[". Authors of [4] think in rare cases link should have multi types, but I think the "rare cases" happen frequently—two object have a high chance to have more than one relationships. Such as a film can be starred, written, directed and produced by Charles Chaplin. Multi-types will save troubles on format and the work load. Attribute is design for data values (which means don't have or don't need to have a link of further information). The data values give the first impressions of objects, such as "Yao Ming is 2.26 meter high", "The lake is about 10meters deep in average" and so on. They normally appear as plain text in wiki, and it is hard to be used for queries and statistic, because machines cannot understand or distinct those from other text. Similar to type link attribute use a value with types in front—[[type:=data value]]. The authors use the signal ":=" instead of "::" to distinct with type link. For example, a sentence "….population is 4,000" will be translated to "[[population: = 4,000]] [4]. The authors of [4] also mentioned a problem – units. For data like population unit might not needed, but what about for area, volume and so on. So there should be a discipline implemented for this kind of data. Conversations between different units should also be performed for future use. Users from different areas have different habits of units using – in China kilometer is preferred when measure the distance between cities while in USA people is more likely to use miles. If the Semantic Wiki site is just for people in a certain area it is not too bad if it doesn't have a unit conversion, but if it faces to several areas the unit conversion is a "Must" requirement. Category is defined as "classify articles according to their content" in [4], and it should be used to manage types and attributes. The kind of type links and attributes a category should be limited, such as u cannot describe a movie with a life period and a city doesn't have a type link called "directed by". A category defines a template of information on a certain object in fact, and will help the auto translations from wiki to semantic wiki, which will discussed in section [3].

## 2.2 New Features

The theories described in section 2.1 make querying and investigating much easier. User can get a list movies stared by Charles Chaplin by searching movies with a type link [[is started by: Charles Champlin]], and list all cities with a population more than 10 million in the world by compare the population of cities with ten million. Of course the jobs of search and comparison are processed by machines. In order to accelerate the processes, tables of type links and attributes should be made. It is much easier for machine to view a table than analyses an article and find type link or attributes. The table should be able to be automatically generated by machines after user finish creating or updating a wiki page.

Another feature of semantic is preventing the not necessary duplications of same piece of information. M.Buffa and F.Gardon gives an example in [6]. In java, "A is a subclass of B" and "A extends B" have the same meaning. In old wiki adding sentence with same meaning to one wiki file cannot be prevent automatically because machines don't really know what the sentence means. In semantic wiki, disciplines can be implemented, such as "there is at most one of those words per page", so the duplications will reduce a lot.

## 3. PRACTICAL APPROACH

The approaches of building a semantic wiki can be divided to two parts, manually and automatically. The forward one can be used to build a new and simple semantic wiki site, because it is not hard for user create or update if a well formatted template is provided, while the backward one is good for converting existing wiki site to semantic ones – manually transform the original wiki files to semantic ones is a waste on money and time.

### 3.1 Manually:

Manually building a semantic wiki seems a good method. Although its cost is high, consider wiki's idea is make user be partner, a big part of the cost is assigned to users. The more users the site has, the less cost is added to a single user. So at least in cost aspect, it is not a big problem. One template is not hard to build for an experienced programmer. For example, a template of movie for the wiki site described at the start of section 2 can be: The template contains most of important information of a movie. Information of a movie can be make up by connected the information by words. Users can use that to update easily as well – just modify the content in the cells, easier than find the content's position in article and modify. So manually building is a good solution? No.

Manually Solution has several problems, or potential problems.

First of all, the example list above is just for a movie wiki site. The target range of that site is quite narrow: movies, peoples related, presented companies, maybe first published countries as well. So the number of templates is quite small. What about a wiki site like wikipeadia, a real encyclopaedia for the world? The templates needed are counted by thousand. It is a hard job to select a suitable template, what even worse is sometime new template need to be created. Creation of new template is hard for normal users.

Second, it increases the complexity of creating a wiki page. As described in section 2, wiki's success is from public attended. The complexity increasing add difficulty for public attend wiki growing. Public's attending is a consciously, voluntary action. The complexity rising will reduce these, and the finally rise the cost of the wiki site.

### 3.2 Automatically

Automatically solution means user input plain text content, and the semantic wiki site will translate it to semantic ones with type links and attributes. It is quite hard to implement, because it is hard to teach the machine how tell the difference between similar expressions and how to get the similarity from different expressions. So far there is no complete solution for this, but I believe this is where the semantic wiki future is. As described in section one, wiki's success is build on public attention. Some people may be experts on a certain area while only have little knowledge on how to use a computer. Wiki gives a way to them to share their knowledge on the net. Reduce cost by rising complexity is not a good idea because it make the wiki step away from the public, which is the base of its success.

F.Wu and D.S.Weld describe their design in "Autonomously Semantifying Wikipedia", which will provide some hint on automatically generating semantic wiki files—they want an approach automatically structuring a large amount of existing data. Some wiki sites have some semantic information, such as wikipeadia info box is used as an abstract. F.Wu and D.S.Weld generate information from info box in [7]. The process is not hard and that's a hint for automatic solutions: some wiki already have structures for that. They also point out even in those "prepared" wiki sites the popularizing rate is not optimize. Those structures are not pre-request for a wiki page, thus a large amount of wiki pages don't have that – in wikipeadia less than 50% of class "U.S Countries" have an info box[7]. Authors worried about duplicated semantic information as well, but use the technology mentioned in section 2.2 can avoid most of these cases. Generating type links and attributes from plain text is a hard job. Basically, it looking for keywords (etc. published, area) in text, and use the information around that to build semantic information. The processes sounds easy, but even define the meaning of "keyword" is a huge project, how to decide the word around is "information" is another difficult problem as well. Authors of [4] make interesting tries called Sentence Classifier. It uses a table to define a certain kind of data values. Table 1 is an example from [4]. Authors admit it target range is very narrow and the efficiency is not high, but it is still a good attempt on automatic generating semantic wiki.

| Feature Description | Example |
|---|---|
| First token of sentence | Hello world |
| In first half of sentence | Hello world |
| In second half of sentence | Hello world |
| Start with capital | Hawaii |
| Start with capital, end with period | Mr. |
| Single capital | A |
| All capital, end with period | CORP. |
| Contains at least one digit | AB3 |
| Made up of two digits | 99 |
| Made up of four digits | 1999 |
| Contains a dollar sign | 20$ |
| Contains an underline symbol | km_square |
| Contains an percentage symbol | 20% |
| Stop word | the; a; of |
| Purely numeric | 1929 |
| Number type | 1932; 1,234; 5.6 |
| Part of Speech tag | |
| Token itself | |
| NP chunking tag | |
| String normalization: | |
| capital to "A", lowercase to "a", | |
| digit to "1", others to "0" | TF − 1 =) AA01 |

| | |
|---|---|
| Part of anchor text | Machine Learning |
| Beginning of anchor text | Machine Learning |
| Previous tokens (window size 5) | |
| Following tokens(window size 5) | |
| Previous token anchored | Machine Learning |
| Next token anchored | Machine Learning |

**Table 1(from page 4 of [7])**

## 4. FUTURE DEVELOPMENT

I read some papers related to semantic wiki. Most of them focus on developing technologies for more powerful functions. But in my opinion, before implement those technologies, the most important thing need to be consider is "how much use complexity is added". As discussed in section one, the ease of use is one of the important features, which attract a lot of users for wiki. That's why this paper interest on automatic generating semantic wikis.

In my opinion, a semantic wiki should have more functions than original wikis, but the input should stay simple—that is, user input a plain text as what they do in wiki, and the plain text will be analyzed, re-structured and stored in semantic way. User does the same amount (maybe a little more) of word load, but produce more meanings. Consider the complexity of the automatic generating tool might have, it is not a good idea run that on the semantic wiki server. The extra work might decrease the server speed a lot [7]. A client side generating tool should be a good idea. Expert users may fix errors or not suitable translations, and normal users get a chance to learn the working style of the tool, which will give them experience on using the tool. After the translation user will upload the translated semantic wiki files to server.

Another direction I can think is the merging. The program described in [7] can gather semantic information from other wikis and use them as self storage in design. With permit a semantic wiki site should be able to generated semantic information from other wiki sites. What the world need is a comprehensive encyclopaedia, which contains nearly all information about the world, not several encyclopaedias which focus on certain area respectively. In that situation, how to keep that efficient will be problem need to be solved.

## 5. CONCLUSIONS

Wiki is a powerful tool for different peoples communicate knowledge on the web. It is quite easy to use and friendly to public, so more and more people use that as a communication platform, and the amount of information stored by wiki increase at the same time. In order to make the information more useful, semantic wiki is introduced. The theory is not hard, wiki save the information with a title-content model, semantic use content tables (attribute table, type link table) instead of content. That will make the data value stored more useful – more queries, investigations can be performed. Building and maintaining a semantic wiki site are not easy jobs, because the operation complexity needs to be considered carefully for users. Manually building and maintaining is a direct way, but it may prevent users with limited computer knowledge attending wiki. Automatic approach is not mature yet, but it should be the future direction of semantic wiki developing, because it inherit wiki's success feature. In conclusion, semantic wiki will make the data value contained in wiki plain text much more valuable.

## 6. ACKNOWLEDGMENTS

Thanks for Gerald Weber's support for this seminar report.

## 7. REFERENCE

[1] Cunningham. W. Ross Mayffeld's notes on Cunningham's Keynote at Wikisym 2005. "Ward Cunningham on theCrucible of Creativity".

http://ross.typepad.com/blog/2005/10/ward_cunningham.html

[2] Wales. J, founder of Wikipedia / Presentation at Wiki Symposium 2005.

http://recentchanges.info/?p=5

[3] Robert Tolksdorf, Elena Paslaru Bontas Simperl Towards Wikis as Semantic Hypermedia

http://delivery.acm.org.ezproxy.auckland.ac.nz/10.1145/1150000/1149470/p79-tolksdorf.pdf?key1=1149470&key2=5889396021&coll=ACM&dl=ACM&CFID=22216675&CFTOKEN=72904275

[4] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, Rudi Studer: Semantic Wikipedia http://delivery.acm.org.ezproxy.auckland.ac.nz/10.1145/1140000/1135863/p585-volkel.pdf?key1=1135863&key2=8762396021&coll=ACM&dl=ACM&CFID=22216675&CFTOKEN=72904275

[5] Eyal Oren, John G. Breslin, Stefan Decker:How semantics make better wikis

http://portal.acm.org.ezproxy.auckland.ac.nz/citation.cfm?id=1135777.1136020&coll=ACM&dl=ACM&CFID=65461779&CFTOKEN=76346162

[6] Michel Buffa, Fabien Gandon: semantic web enabled technologies in Wiki http://delivery.acm.org.ezproxy.auckland.ac.nz/10.1145/1150000/1149469/p69-buffa.pdf?key1=1149469&key2=0120396021&coll=ACM&dl=ACM&CFID=22216675&CFTOKEN=72904275

[7] Fei Wu, Daniel S. Weld:Autonomously Semantifying Wikipedia http://portal.acm.org.ezproxy.auckland.ac.nz/citation.cfm?id=1321440.1321449&coll=ACM&dl=ACM&CFID=65461779&CFTOKEN=76346162