

Wikipedia Mining

Wikipedia as a Corpus for Knowledge Extraction

Kotaro Nakayama, Minghua Pei, Maike Erdmann, Masahiro Ito,
Masumi Shirakawa, Takahiro Hara and Shojiro Nishio

Dept. of Multimedia Eng., Graduate School of Information Science and Technology
Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
TEL: +81-6-6879-4513 FAX: +81-6-6879-4514
{nakayama.kotaro, hai.meika, erdmann.maike, ito.masahiro,
shirakawa.masumi, hara, nishio}@ist.osaka-u.ac.jp

Abstract. Wikipedia, a collaborative Wiki-based encyclopedia, has become a huge phenomenon among Internet users. It covers a huge number of concepts of various fields such as Arts, Geography, History, Science, Sports and Games. As a corpus for knowledge extraction, Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, word sense disambiguation based on URL and brief anchor texts. Because of these characteristics, Wikipedia has become a promising corpus and a big frontier for researchers. A considerable number of researches on Wikipedia Mining such as semantic relatedness measurement, bilingual dictionary construction, and ontology construction have been conducted. In this paper, we take a comprehensive, panoramic view of Wikipedia as a Web corpus since almost all previous researches are just exploiting parts of the Wikipedia characteristics. The contribution of this paper is triple-sum. First, we unveil the characteristics of Wikipedia as a corpus for knowledge extraction in detail. In particular, we describe the importance of anchor texts with special emphasis since it is helpful information for both disambiguation and synonym extraction. Second, we introduce some of our Wikipedia mining researches as well as researches conducted by other researches in order to prove the worth of Wikipedia. Finally, we discuss possible directions of Wikipedia research.

1 Introduction

Wikipedia has become an invaluable corpus for research in various areas such as AI, NLP, Web mining and Semantic Web since it is a database storing all human knowledge that covers a huge number of concepts of various fields. It covers quite large area of topics from general to domain-specific. As a corpus for knowledge extraction, Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, word sense disambiguation based on URL and brief anchor texts. Furthermore, it supports over 250 languages and sentences are well structured. The fact that these characteristics can be used effectively to extract valuable knowledge from Wikipedia is strongly confirmed by a number of researches on Wikipedia Mining [1–6].

Wikipedia research can be categorized by the purpose; semantic relatedness measurement (association thesaurus construction), semantic relation extraction (ontology construction), bilingual dictionary extraction etc. In the middle of 2005, we launched *Wikipedia Lab.*, a special interest research group for Wikipedia mining to investigate the dynamics and capability of Wikipedia. Until now, we have already conducted various Wikipedia mining researches such as the construction of association thesauri, Web ontologies and bilingual dictionaries. By these activities, our conviction that Wikipedia is an invaluable corpus has been confirmed strongly.

The purpose of this paper is to share the knowledge that we acquired in the research activities and take a comprehensive, panoramic view of Wikipedia as a Web corpus aiming to make a compass for Wikipedia researchers. In order to do this, in this paper, we describe the characteristics of Wikipedia, introduce current researches and discuss future directions. The rest of this paper is organized as follows. In section 2, we unveil the characteristics of Wikipedia as a corpus for knowledge extraction with detailed statistics and discussion. In particular, we describe the importance of anchor texts with special emphasis since it is helpful information for both disambiguation and synonym extraction. Second, in section 3, we introduce some of our Wikipedia mining researches with concrete results as well as researches conducted by other researches in order to prove the capability of Wikipedia Mining. We discuss the possible directions of Wikipedia researches in section 4. Finally, we draw a conclusion in section 5.

2 Characteristics of Wikipedia

As we mentioned before, as a corpus for knowledge extraction, Wikipedia has various impressive characteristics such as live updates, word sense disambiguation by URL, the dense link structure and brief anchor texts [7]. In this section, we describe these characteristics in detail.

2.1 Disambiguation by URL

Word sense disambiguation by URL is one of the most notable characteristics of Wikipedia. Ordinary dictionaries have indexes to find the concepts the user wants to look up. However, several concepts are put into one index in most cases. This means that ambiguous terms are listed in one article. This is no problem for humans because it is human readable, but it is not machine understandable.

For example, if a sentence “Golden delicious is a kind of apple” exists in an article in a dictionary, humans can immediately understand that “apple” means a fruit. However, it is difficult to analyze for a machine because “apple” is an ambiguous term and there is no identification information for it. To make this sentence machine understandable, we need some identifier.

In Wikipedia, almost every page (article) corresponds to exactly one concept and has its own URL respectively. For example, the concept apple as a fruit has a Web page and an own URL “<http://en.wikipedia.org/wiki/Apple>”. Further, the computer company Apple also has an own URL “http://en.wikipedia.org/wiki/Apple_Inc.,” so these concepts are semantically separated. This means that it is possible to analyze term relations avoiding ambiguous term problems or context problems.

2.2 Brief Anchor Texts

The *anchor text* is also valuable information. An anchor text is the text part of a link that is shown in the Web page. For instance, the text part “Apple” is the anchor text in the example below.

```
<a href="http://en.wikipedia.com/wiki/Apple_Computer">
Apple
</a>
```

An anchor text is often a summary of the linked page, thus it can be used for various purposes such as Web page categorization[8]. Anchor texts in Wikipedia have a quite brief and simple form compared with those of ordinary Web sites. Anchor texts in ordinary Web corpora often contain wordy information like “Click here for more detailed information.” Sometimes, the anchor text does not contain any important information about the linked page which is one of the common causes of accuracy problems on Web mining based thesaurus construction. As opposed to that, anchor texts in Wikipedia are refined very well.

Among the authors of Wikipedia, it is a common practice to use the title of an article for the anchor text, but users also have the possibility to give other anchor texts to an article. To create a hyperlink to another page in Wikipedia, the special tag “[[...]]” is used. For example, the sentence “Kyoto is a [[city]]” contains a hyperlink to the page (<http://www.wikip-edia.org/wiki/city>) which describes the general meaning of the word city. Editors also can provide a anchor text for the hyperlink using a piped text (“|”).¹ For instance, a link “[[Mobile phone | Cell phone]]” has the link text “Cell phone” and it refers to the article “Mobile phone.” This feature makes another important characteristic; the “variety of anchor texts,” which can be used for both disambiguation and synonym extraction. In the Wikipedia researches, surprisingly few studies have so far been made at anchor text analysis despite of the importance. It seems likely that the technical difficulty of anchor text extraction is the reason why only few attempts have been made at anchor text analysis. Since Wikipedia dump data does not contain anchor text information, our research group developed a enhanced parser for MediaWiki and extracted all anchor texts in all pages in Wikipedia.

2.3 Live Updates

The content management work flow of ordinary dictionaries made by human effort is a top-down approach in most cases. The top-down approach is to the advantage of the quality, but to the disadvantage of the topic coverage. This means that general concepts will be covered first, and domain specific terms/new concepts will be covered later (or never). For instance, almost all paper-based dictionaries have no entry for iPhone even though Wikipedia has several entries for e.g. iPhone with detailed information and pictures.

The work flow of Wikipedia is based on a bottom-up approach. Since Wikipedia is Wiki[9] based, it allows users to edit articles easily and timely. This feature

¹ For further information about Wikipedia syntax, see http://en.wikipedia.org/wiki/Wikipedia:How_to_edit_a_page

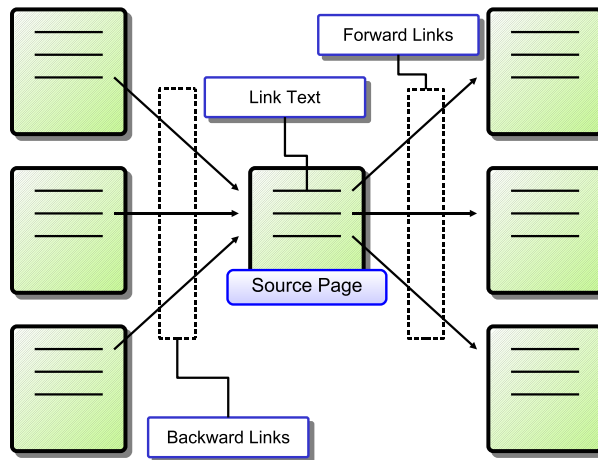


Fig. 1. Various information on hyperlinks.

leads to several advantages; wide-range concept coverage, new concept coverage, and collaborative modification. As an example, after the announcement of a latest product, an article on the product with detailed information is usually uploaded a lot faster than on ordinary paper-based dictionaries. One of the most difficult issues of thesaurus construction is the coverage of new terms, but this characteristic shows that Wikipedia has the potential to overcome this problem.

After all, the ease of use was the dominant factor of success in wide-range topic coverage. Wikipedia allows users to edit content via Web browsers. Since authorities on specific topics are not always good at using complicated computer systems, this critical feature helped to gather so many contributors and to cover wide-range topics.

2.4 Link Structure of Wikipedia

The dense link structure is one of the most interesting characteristics of Wikipedia. “Dense” means that it has a lot of “inner links,” links from pages in Wikipedia to other pages in Wikipedia. This means that articles are strongly connected by many hyperlinks. By analyzing the link structure, we can extract various information such as topic locality[10], site topology, and summary information. Topic locality is the law that Web pages which are sharing the same links have more topically similar contents than pages which are not sharing links. We believe that Wikipedia has topic locality and the connectivities among articles are much stronger than on ordinary Web sites because of the dense link structure.

Let us show statistics of link structure analysis for Wikipedia which we investigated. Figure 2 shows the distribution of both backward links and forward links. Both of them have typical Zipf distribution, containing a few nodes that have a very high degree links and many with low degree links. 196 pages have

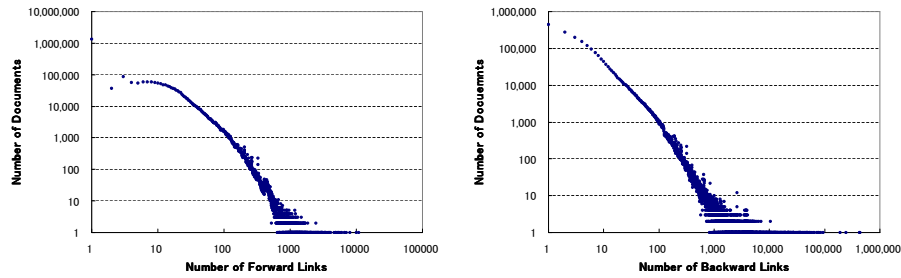


Fig. 2. Zipf distribution on the Wikipedia link structure.

more than 10,000 backward links/page, 3,198 pages have more than 1,000 backward links/page and 67,515 pages have more than 100 backward links/page.

From the English Wikipedia, 49,980,910 forward links (excluding redirect links) were extracted from 1,686,960 pages (excluding redirect, image, category pages). This means a page in Wikipedia has 29.62 forward links on average. Further, 2,531 pages have more than 500 forward links/page and 94,932 pages have more than 100 forward links/page. It can be concluded, from the statistics has been shown above, Wikipedia has a very dense link structure.

The statistics shows that we need to consider the characteristics to design algorithms for analyzing the Wikipedia link structure.

2.5 Link Types

Wikipedia contains several types of links such as interlanguage links, category links and redirect links. All of them have impressive characteristics and are useful information.

An *interlanguage link* is a link between two articles in different languages. We assume that in most cases, the titles of two articles connected by an interlanguage link are translations of each other. Currently, Wikipedia supports over 250 languages including major languages, minor languages and auxiliary languages (such as Esperanto), and has a dense link structure among major languages. For instance, there are 124,357 interlanguage links (Sept. 2006) from the English Wikipedia to the Japanese Wikipedia. This means that a considerable number of translation term pairs can be extracted. As a comparison, EDICT, one of the largest manually created English-Japanese dictionaries, has about 110,000 entries.

A *category link* is a link which is used to define taxonomic relation between an article and a category in Wikipedia. In other words, a category link is used to define what category the article belongs to. Furthermore, category links are also used to define relations between categories. It is widely known that Wikipedia's category tree is well organized (totally 6,979,355 category links exist in Sept. 2006) and it is already used in various researches such as semantic relatedness measurement and semantic relation extraction. However, what has to be noticed is that the category tree is not an ontology, but just a taxonomy. Therefore,

we must conduct deeper analysis for determining the explicit relation type of a category link.

Redirect pages in Wikipedia are pages containing no content but a link to another article (target page) in order to facilitate the access to Wikipedia content. When a user accesses a redirect page, he will automatically be redirected to the target page. Redirect pages are usually strongly related to the concept of the target page. They often indicate synonym terms, but can also be abbreviations, more scientific or more common terms, frequent misspellings or alternative spellings etc. The notable fact is that about one third the pages in Wikipedia (totally 1,267,162 pages in Sept. 2006) are redirect pages.

3 Wikipedia Mining Researches

In this section, we introduce a number of Wikipedia researches with concrete examples of results to show the capability of the methods.

3.1 Wikipedia Thesaurus

Wikipedia Thesaurus [3] is an association thesaurus constructed by mining the Wikipedia link structure. An association thesaurus is a set of concepts and relations among the concepts. In Wikipedia researches [1–4], it is strongly proved that Wikipedia is suitable for this aim. WikiRelate [4] is one of the pioneers in this research area. The algorithm finds the shortest path between the categories which the concepts belong to in a category graph.

Besides, we proposed a scalable link structure mining method named *pfibf* (Path Frequency - Inversed Backward link Frequency) to extract a huge scale association thesaurus in a previous research [3]. *pfibf* measures the relatedness between two articles (concepts) by analyzing the links among them. In this method, the relativity between any pair of articles (v_i, v_j) is assumed to be strongly affected by the following two factors:

- the number of paths from article v_i to v_j ,
- the length of each path from article v_i to v_j .

The relativity is strong if there are many paths (sharing of many intermediate articles) between two articles. In addition, the relativity is affected by the path length. In other words, if the articles are placed closely together in the concept graph and sharing hyperlinks to articles, the relativity is estimated to be higher than further ones.

The number of backward links of an article is also estimated as a factor of relativity because general/popular articles have a lot of backward links and these articles easily have high relativity to many articles. Therefore, we must consider the inversed backward link frequency *ibf* in addition to the two factors above. The relatedness between two articles becomes stronger if there are many paths between them. In addition, the relativity becomes stronger according to the path length.

Therefore, if all paths from v_i to v_j are given as $T = \{t_1, t_2, \dots, t_n\}$, the relativity *pf* (path frequency) between them can be expressed as follows:

Table 1. Sample of queries related terms extracted by *pfibf*.

Query	Extracted association terms		
Sports	Basketball	Baseball	Volleyball
Microsoft	MS Windows	OS	MS Office
Apple Comp.	Macintosh	Mac OS X	iPod
iPod	Apple Comp.	iPod mini	iTunes
Book	Library	Diamond Sutra	Printing
Google	Search engine	PageRank	Google search
Horse	Rodeo	Cowboy	Horse-racing
Film	Actor	Television	United States
DNA	RNA	Protein	Genetics
Canada	Ontario	Quebec	Toronto

$$pfibf(v_i, v_j) = pf(v_i, v_j) \cdot ibf(v_j), \quad (1)$$

$$pf(v_i, v_j) = \sum_{k=1}^n \frac{1}{d(|t_k|)}, \quad (2)$$

$$ibf(v_j) = \log \frac{N}{bf(v_j)}. \quad (3)$$

$d()$ denotes a function which increases the value according to the length of path t_k . A monotonically increasing function such as a logarithm function can be used for $d()$. N denotes the total number of articles and $bf(v_j)$ denotes the number of backward links of v_j . This means a page which shares hyperlinks with a specific page but not with other pages, has a high *pfibf*.

However, counting all paths between all pairs of articles in a huge graph is a computational resource consuming work. Therefore, we proposed an efficient data structure named “Dual binary tree” (DBT) and a multiplication algorithm for the DBT[3].

Table 1 shows an example of an association thesaurus constructed by *pfibf*. As we can see, the extracted terms are quite related to the queries. In fact, we confirmed that the accuracy is much better than traditional methods such as co-occurrence analysis and TF-IDF. We currently extracted over 243 million association relations among 3.8 million Wikipedia concepts.

3.2 Disambiguation by Anchor Text

As we mentioned before, anchor texts contain valuable information although only few attempts have been made so far. Anchor texts are useful for both disambiguation and synonym extraction. In this subsection, we describe how anchor texts can be used for word sense disambiguation first.

Ambiguous terms have the same expression but several meanings. For instance, “Apple” can mean a computer company or a fruit. After a number of experiments, we realized that anchor texts are helpful information for ambiguous term detection and disambiguation because we easily detect whether a word is

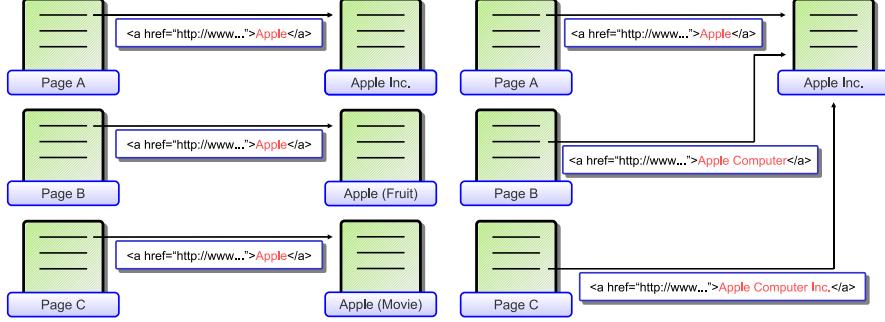


Fig. 3. Ambiguous/synonym term detection by backward link analysis.

ambiguous by analyzing all links on Wikipedia. If same expression is widely used on links pointing to different articles, the term might be an ambiguous term. Figure 3 (left side) illustrates our idea. For instance, for a term “Apple,” apple (a fruit) and Apple computer (a company) both have the anchor text “apple” in their backward links. This means that there are two major meanings for a term “apple.” In a previous research [7], we proposed a concept specification method. For a given natural language query string q , the concept specification method CS is defined as follows:

$$CS(v_i, q) = \frac{Cnt(B_{v_i}|q)}{\sum_{v_j \in V} Cnt(B_{v_j}|q)}. \quad (4)$$

$Cnt(B_{v_i}|q)$ is a function that returns the number of backward links of page v_i having the anchor text q . According to the result of several experiments, this strategy has proved to be effective. For instance, for the query “Apple” as q , the CS value scored 0.40 for apple as a fruit, 0.44 for Apple Computer as a computer company, and 0.09 for Apple Records as a record label. This means that the term “Apple” is widely used in roughly two meanings; a fruit and a computer company. One more example; “UFO” as q and “Unidentified_flying_object” as v_i , the CS value scored 0.65. This means that the query “UFO” is estimated to be equivalent to “Unidentified_flying_object” by 65% of possibility. By providing all possible candidates to the user and letting the user choose one of them, the meaning of a term can be specified.

3.3 Synonym Extraction from Anchor Texts

We discovered that we can extract synonyms by analyzing the anchor text of an article in previous research [7]. A synonym is a term that has one meaning but various expressions. For instance, “Apple Computer” is sometimes referred to as just “Apple.” Since backward links of a Web page have a “variety of backward anchor texts,” this variety can be used to extract synonyms of a concept (article). Figure 3 (right side) shows an example of the variety of backward anchor texts. “Apple Computer” is sometimes just written as “Apple”, but sometimes written

Table 2. Synonym extraction by anchor text analysis.

Concept	Synonyms
Apple Computer	'Apple' (736), 'Apple Computer, Inc.' (41), 'Apple Computers' (17)
Macintosh	'Apple Macintosh' (1,191), 'Mac' (301), 'Macs', 30
Microsoft Windows	'Windows' (4,442), 'WIN' (121), 'MS Windows' (98)
International Organization for Standardization	'ISO' (1,026), 'international standard' (4), 'ISOs' (3)
Mobile phone	'mobile phones' (625), 'cell phone' (275), 'Mobile' (238)
United Kingdom	'United Kingdom' (50,195), 'British' (28,366), 'UK' (24,300)

(): Number of backward links (Anchor texts corresponding to the title of an article are excluded).

as “Apple Computer, Inc,” “Apple Computers,” etc. Table 2 shows a number of examples of randomly chosen synonym terms.

The article “Apple Computer” has 1,191 backward links with the anchor text “Apple Macintosh” and 301 backward links with the anchor text “Mac.” This shows that both words are typical synonyms for the concept “Apple Computer.” Statistical data unveiled that backward anchor texts analysis can extract high quality synonyms by specifying a threshold to filter noisy data such as “international standard” and “ISOs” for ISO.

Synonyms are quite helpful information to detect whether sentences are describing the same subject. In other words, the information is needed for co-reference resolution. For example, there is an article about “United Kingdom” in Wikipedia and it contains the word “UK” many times. However, without the fact that “UK” is a synonym of “United Kingdom,” it cannot extract many relations on the topic. So, we use the extracted synonyms.

3.4 Wikipedia Ontology

Wikipedia Ontology is a huge scale Web ontology automatically constructed from Wikipedia. We proposed a consistent approach of semantic relation extraction from Wikipedia. To extract inferable semantic relations with explicit relation types, we need to analyze not only the link structure but also texts in Wikipedia. However, parsing Wikipedia text is not a trivial work since Wikipedia is written by natural language with Wiki tags (E. g. quotations, hyperlinks, tables and HTML tags). Therefore, we need to develop processors including trimmer, chunker and POS tree analysis by ourselves. The method consists of three sub-processes highly optimized for Wikipedia mining; 1) fast preprocessing (trimming, chunking, and partial tagging), 2) POS tag tree analysis, and 3) mainstay extraction. Table 3 shows examples of the result of our proposed method.

Basically, the proposed method extracts semantic relations by parsing texts and analyzing the structure tree generated by a POS parser. Currently, we support three patterns for POS tag analysis; definitive sentence pattern (e.g. “is-a”), subordinating pattern (E. g. “is-a-part-of”) and passive pattern (E. g. “was-born-in”).

Table 3. Examples of extracted explicit relations by ISP.

Subject	Predicate	Object
Odonata	is an order of	Insect
Clarence Thomas	was born in	Pin Point, Georgia
Dayton, Ohio	is situated within	Miami Valley
Germany	is bordered on	Belgium
Germany	is bordered on	Netherlands
Mahatma Gandhi	founded	Natal Indian Congress
Mahatma Gandhi	established	Ashram
Rice	has	Leaf
Rice	is cooked by	Boiling
Rice	is cooked by	Steaming

Semantic Wikipedia [11] is one of the pioneers in this research area. Semantic Wikipedia is an extension of Wikipedia which allows editors to add semantic relations manually. Another interesting approach is to use Wikipedia’s category tree as an ontology [12, 13, 6]. Wikipedia categories are a promising resource for ontology construction, but categories can not be used as an ontology since they are just a taxonomy and do not provide explicit relation types among concepts. In our Wikipedia Ontology project, in contrast to these approaches, we developed a full-automated consistent approach for semantic relation extraction by mining Wikipedia article texts.

3.5 Wikipedia API

Wikipedia API is an XML Web service that enables users to add the capability of Wikipedia Mining methods to their own applications. We currently provide disambiguation, synonym extraction and association thesaurus look up functions for developers. The thesaurus is stored in a database (MySQL 5.0) and indexed to provide practical APIs for other applications such as information retrieval and text summarization. The APIs are provided by XML Web services to achieve high interoperability among different environments. *Wikipedia Ontology* will also be provided as an XML Web service in near future.

We developed a thesaurus browsing system by using the APIs to prove the capability of this approach and the accuracy of our association thesaurus. Figure 4 shows the architecture of our system and a screen shot of the thesaurus browsing system.

The concept search engine provides SKOS[14] (an application of RDF) representation function. SKOS is a developing specification and standard to support knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the Semantic Web. SKOS supports several basic relations between two concepts such as “broader,” “narrower” and just “related.” Further, since *pfibf* extracts relatedness between two concepts, we extended SKOS in order to express the relatedness. We defined an extension for SKOS called the “Wikipedia Thesaurus Vocabulary (WTV).” Currently, WTV supports some simple relations not in SKOS such as relatedness. The URLs for concepts in the Wikipedia Thesaurus correspond to the URLs of

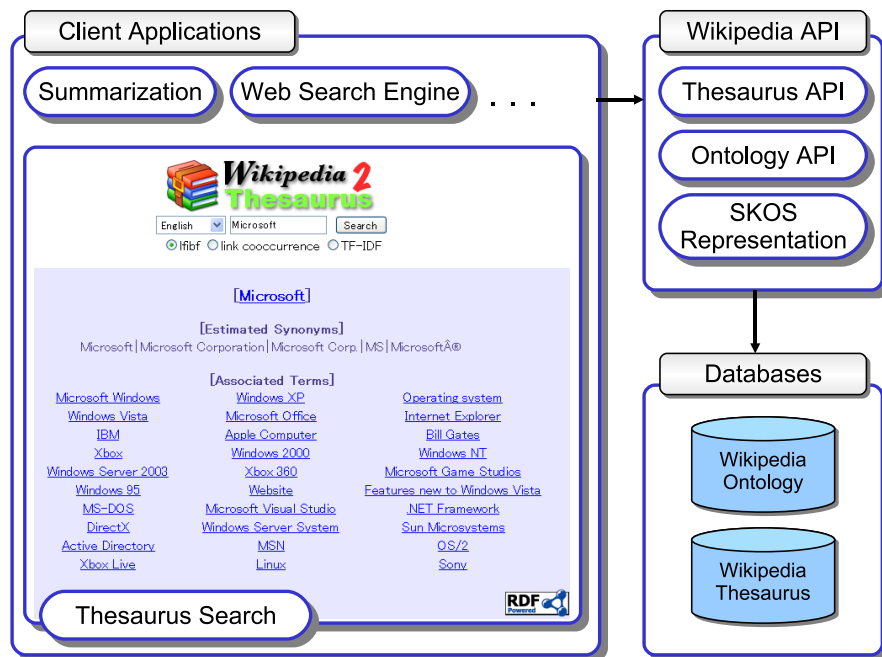


Fig. 4. Search engine for Wikipedia Thesaurus and architecture

Wikipedia articles. For example, “<http://wikipedia-lab.org/concept/Music>” in the thesaurus corresponds to the article “<http://en.wikipedia.org/wiki/Music>” in Wikipedia. Figure 5 shows an example of a SKOS representation of “Music.”

Wikipedia Thesaurus Visualizer (Figure 6) is another application of Wikipedia API. It is a client/server visualization system for Wikipedia Thesaurus and uses the Wikipedia API as a backside database to draw a concept network in n-hop range. It draws a concept network by invoking the API asynchronously since getting neighbors in n-hop range is time consuming and would lead to delay in most cases.

3.6 Bilingual Dictionary Construction

As we mentioned before, since Wikipedia supports over 250 languages and has a huge number of interlanguage links among articles in different languages, it has become an impressive corpus for bilingual terminology construction. However, interlanguage link are basically one by one mapping between two language articles. This means that we can extract only one translation for each concept. To achieve high coverage of translation, in our previous research [15], we proposed an enhancement method using three types of information; interlanguage links, redirect links and anchor texts. The flow is described as follows.

At first, we create a baseline dictionary from Wikipedia by extracting all translation candidates from interlanguage links. For a term to be translated, a

```

- <rdf:RDF>
- <skos:Concept rdf:about="http://wikipedia-lab.org/concepts/Music" rdfs:seeAlso="http://en.wikipedia.org/wiki/Music">
  <skos:prefLabel>Music</skos:prefLabel>
  <skos:altLabel>musical</skos:altLabel>
  <skos:altLabel>musician</skos:altLabel>
  <skos:altLabel>musicians</skos:altLabel>
  <skos:altLabel>genres</skos:altLabel>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Performance" wtv:relatedness="0.0579156"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Music_venue" wtv:relatedness="0.0564693"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Publication" wtv:relatedness="0.0230024"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Dance" wtv:relatedness="0.0170662"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Organization" wtv:relatedness="0.0111082"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Art" wtv:relatedness="0.0104583"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Drama" wtv:relatedness="0.00829775"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Theatre" wtv:relatedness="0.00791607"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Philosophy" wtv:relatedness="0.00704391"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Poetry" wtv:relatedness="0.00689856"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Music_theory" wtv:relatedness="0.00689445"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Pitch_(music)" wtv:relatedness="0.00686802"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Painting" wtv:relatedness="0.00658941"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Literature" wtv:relatedness="0.00651561"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Theater" wtv:relatedness="0.00628334"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Piano" wtv:relatedness="0.00589881"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Lyrics" wtv:relatedness="0.00568608"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Melody" wtv:relatedness="0.00563164"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Entertainment" wtv:relatedness="0.00556938"/>
  <skos:related rdf:resource="http://wikipedia-lab.org/concepts/Composer" wtv:relatedness="0.00548984"/>

```

Fig. 5. Sample of SKOS (RDF) representation

Wikipedia source page sp is extracted if its title s is equivalent to that term. In cases where the term is equivalent to the title of a redirect page, the corresponding target page is used as sp and its title as s . After that, in case sp has an interlanguage link to a page tp in the target language, the title t of tp will be chosen as the translation.

In the second step, we enhance the number of translations by using redirect pages. The idea is to enhance the dictionary with the set of redirect page titles R of all redirect pages of page tp . As mentioned before, not all redirect pages are suitable translations. Therefore, we want to assign a score to all extracted translation candidates and filter doubtful terms through a threshold. We found out experimentally that the number of backward links of a page can be used to estimate the accuracy of a translation candidate, because redirect pages where the title is wrong or semantically not related to the title of the target page usually have a small number of backward links. This approach has already proved effective in creating the Wikipedia Thesaurus [3].

We calculate the score of a redirect page title r of a redirect page rp by comparing the number of backward links of rp to the sum of backward links of tp and all its redirect pages.

The score s_{rp} is hence defined by the formula:

$$s_{rp} = \frac{|\text{Backward links of } rp|}{|\text{Backward links of } tp \text{ and of all redirect pages of } tp|}. \quad (5)$$

We can calculate the score of the target page title t in an analogous manner. Usually, redirect pages have much less backward links than target pages.

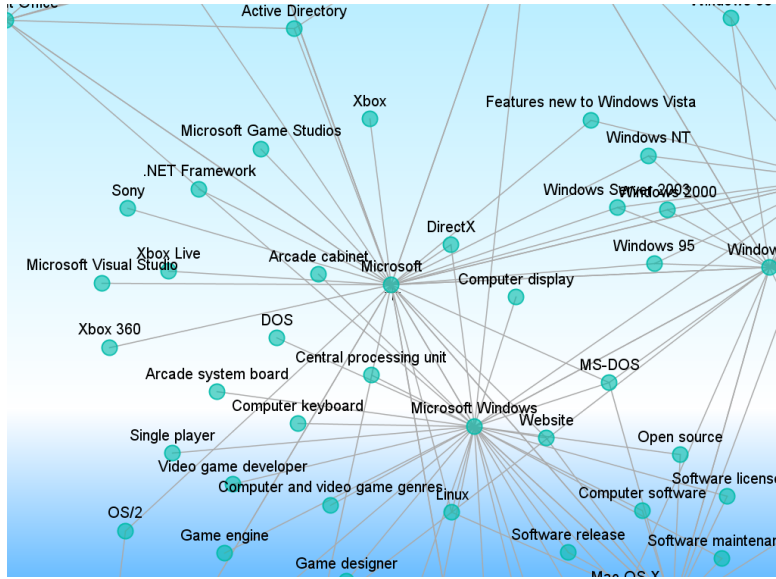


Fig. 6. Wikipedia Thesaurus Visualization

However, redirect pages with more backward links than the corresponding target page also exist, indicating that the redirect page title is a good translation candidate, potentially even better than the target page title.

In the third step, we enhance the translation by using the set of anchor texts LT of all inner language backward links of tp . Like for the RP method, we filter unsuitable translations extracted by the LT method by setting a threshold. We calculate the score s_{lt} of an anchor text lt by comparing the number of backward links of tp containing the anchor text lt to the total number of backward links of tp :

$$s_{lt} = \frac{|\text{Backward links of } tp \text{ with anchor text } lt|}{|\text{Backward links of } tp|}. \quad (6)$$

We conducted a detailed evaluation to investigate the performance of our proposed method compared with the extraction of bilingual terminology from parallel corpora, a state of the art in this research area. In the experiments, our conviction that Wikipedia is an invaluable resource for bilingual dictionary extraction and that redirect pages and anchor texts are helpful to enhance a dictionary constructed from interlanguage links has been confirmed. Our methods are very useful for specialized domain-specific terms (terms usually not contained in manually constructed dictionaries), because accuracy and coverage are much better than that of the parallel corpus approach and also better than the baseline dictionary created from interlanguage links only.

4 Future Works

From the success of current researches on Wikipedia, there is no doubt that Wikipedia is an invaluable corpus for knowledge extraction and can be used for various research areas such as AI, NLP, Web mining and Semantic Web. However, there are technical issues such as reliability of information, scalability and accuracy/coverage of mining method. This means that Wikipedia is a big frontier for researchers and there are many research areas we can contribute to.

What we are mainly tackling now is the *knowledge structuring* project on Wikipedia. *Wikipedia Ontology* is a part of knowledge structuring, but we are trying to extract more advanced relations among concepts. In this project, “integration” of individual Wikipedia researches will be a key solution. For instance, Wikipedia Thesaurus, an association thesaurus, can be used as a basic infrastructure for various researches including ontology construction because relatedness is a helpful information to extract important relations. Furthermore, like YAGO [6], integration with existing ontologies such as WordNet also will be a promising approach since Wikipedia has great coverage for domain-specific terms and existing ontologies cover general terms.

5 Conclusion

In this paper, we unveiled the characteristics of Wikipedia in detail and introduced a number of Wikipedia researches with concrete results to show the capability of Wikipedia Mining. Finally, we explained several technical issues and discussed the future vision of Wikipedia research. Our achievements are available on the WWW and can be accessed from following URLs.

- Wikipedia Lab.
<http://wikipedia-lab.org>
- Wikipedia Thesaurus
<http://wikipedia-lab.org:8080/WikipediaThesaurusV2>
- Wikipedia Ontology
<http://wikipedia-lab.org:8080/WikipediaOntology>
- Wikipedia API
http://wikipedia-lab.org/en/index.php/Wikipedia_API
- Wikipedia Bilingual Dictionary
<http://wikipedia-lab.org:8080/WikipediaBilingualDictionary>

We hope that this paper will be a helpful compass for researchers who want to conduct research on Wikipedia Mining. Furthermore, we believe that conducting research on Wikipedia will be a helpful activity for the Wikipedia community. For instance, showing reliability of information or relatedness among concepts are quite helpful information for both readers and contributors. We hope our activities will be the beginning of an eco-system among Wikipedia readers, contributors and researchers.

6 Acknowledgment

This research was supported in part by Grant-in-Aid on Priority Areas (18049050), and by the Microsoft Research IJARC Core Project.

References

1. E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pp. 1606–1611, 2007.
2. D. Milne, O. Medelyan, and I. H. Witten, "Mining domain-specific thesauri from wikipedia: A case study," in *Proc. of ACM International Conference on Web Intelligence (WI'06)*, pp. 442–448, 2006.
3. K. Nakayama, T. Hara, and S. Nishio, "Wikipedia mining for an association web thesaurus construction," in *Proc. of IEEE International Conference on Web Information Systems Engineering (WISE 2007)*, pp. 322–334, 2007.
4. M. Strube and S. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," in *Proc. of National Conference on Artificial Intelligence (AAAI-06)*, pp. 1419–1424, July 2006.
5. D. P. T. Nguyen, Y. Matsuo, and M. Ishizuka, "Relation extraction from wikipedia using subtree mining," in *Proc. of National Conference on Artificial Intelligence (AAAI-07)*, pp. 1414–1420, 2007.
6. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*, (New York, NY, USA), pp. 697–706, ACM, 2007.
7. K. Nakayama, T. Hara, and S. Nishio, "A thesaurus construction method from large scale web dictionaries," in *Proc. of IEEE International Conference on Advanced Information Networking and Applications (AINA 2007)*, pp. 932–939, 2007.
8. S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks," *SIGMOD Rec.*, vol. 27, no. 2, pp. 307–318, 1998.
9. W. C. Bo Leuf, *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley, 2001.
10. B. D. Davison, "Topical locality in the web," *Proc. of the ACM SIGIR*, pp. 272–279, 2000.
11. M. Völkel, M. Krötzsch, D. Vrandečić, H. Haller, and R. Studer, "Semantic wikipedia," in *Proc. of International Conference on World Wide Web (WWW 2006)*, pp. 585–594, 2006.
12. S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou, "Extracting semantics relationships between wikipedia categories," in *Proc. of Workshop on Semantic Wikis (SemWiki 2006)*, 2006.
13. D. N. Milne, O. Medelyan, and I. H. Witten, "Mining domain-specific thesauri from wikipedia: A case study," in *Web Intelligence*, pp. 442–448, 2006.
14. World Wide Web Consortium, "Simple knowledge organisation systems (skos)," <http://www.w3.org/2004/02/skos/>, 2004.
15. M. Erdmann, K. Nakayama, T. Hara, and S. Nishio, "An approach for extracting bilingual terminology from wikipedia," in *Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA)*, March 2008.