Extracting Semantic Relationships between Wikipedia Categories

Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou

L3S Research Centre, University of Hannover, Expo Plaza 1, 30539, Hannover, Germany,
{chernov, iofciu, zhou, nejdl}@l3s.de

Abstract. The Wikipedia is the largest online collaborative knowledge sharing system, a free encyclopedia. Built upon traditional wiki architectures, its search capabilities are limited to title and full-text search. We suggest that semantic information can be extracted from Wikipedia by analyzing the links between categories. The results can be used for building a semantic schema for Wikipedia which could improve its search capabilities and provide contributors with meaningful suggestions for editing the Wikipedia pages. We analyze relevant measures for inferring the semantic relationships between page categories of Wikipedia. Experimental results show that Connectivity Ratio positively correlates with the semantic connection strength.

1 Introduction

The Wikipedia [1] is a freely accessible Web encyclopedia. The Wikipedia project started in 2001 as a complement to the expert-written Nupedia and it is currently run by the Wikipedia Foundation. There are Wikipedia versions in 200 languages, with more than 3,700,000 articles and 760,000 registered users. An especially interesting aspect of Wikipedia is the categorization and linkage within its content. Pages in Wikipedia are explicitly assigned to one or more *Categories*. Categories should represent major topics and their main use within Wikipedia is in finding useful information. There are two types of categories. The first type is used for classification of pages with respect to topics. They can have hierarchical structure, for example the page can be assigned to the category *Science* or one of its subcategories like *Biology* and *Geography*. The second type of categories is *Lists*, they usually contain links to instances of some concept, for example *List of Asian Countries* points to 54 Asian countries. There also exist numerous links between pages. While most of them are created to provide efficient navigation over the Wikipedia contents, they also represent some semantic relationships between pages or categories.

Like in most of the wikis, the search capabilities on Wikipedia are limited to traditional full-text search, while search could benefit from the rich Wikipedia semantics and may allow complex searches like *find Countries which had Democratic Non-Violent Revolutions*. Using categories as a loose database schema, we can enrich Wikipedia search capabilities with such complex query types. Wikipedia categories could be organized in a graph, where the nodes are categories and the edges are hyperlinks. For example, if some page from the category "Countries" points to a page from the category "Capitals" we can establish a connection "Countries to Capitals". However, not all hyperlinks in Wikipedia are semantically significant such that they can be used to facilitate search. The problem is how to distinguish strong semantic relationships from irregular and navigational links.

In this paper we propose two measures for automatic filtering of strong semantic connections between Wikipedia categories. One measure is the number of links between pages in two categories, and the other is Connectivity Ratio. They can be applied to inlinks or outlinks separately. For evaluation, we apply these measures to the English Wikipedia and perform user study to assess how semantically strong the extracted relationships are. We observe that both number of links and Connectivity Ratio correlates with semantic connection strength. It supports our hypothesis, while much more experiments are needed to achieve a convincing evaluation.

The rest of the paper is organized as follows. The related work is given in Section 2. In Section 3 we describe in detail the problem of discovering strong semantic relationships between categories and the possible use of semantic scheme in Wikipedia. Later, in Section 4 we describe our analysis of factors, relevant for discovering semantic links and present our experiments in Section 5. We conclude and outline future research directions in Section 6.

2 Related Work

The idea to bring semantics into Wikipedia is not new, several studies on this topic have been carried out in the last few years.

The semantic relationships in Wikipedia were discussed in [10]. The authors considered the use of link types for search and reasoning and its computational feasibility. Its distinctive feature is the incorporation of semantic information directly into wiki pages. Later, the semantic links proposal was extended in [12] to the Semantic Wikipedia vision. According to this model, the pages annotations should contain the following key elements: categories, typed links, and attributes. Typed links in form of *is capital of* are introduced via markup extension [[is capital of::England]], each link can be assigned multiple types. They also proposed the usage of semantic templates, based on the existing Wikipedia templates. We follow this approach, but concentrate on automatic extraction instead of manual link assignment. Also, our goal is to enable better search on Wikipedia, but not to provide means for full-fledged reasoning. So we can tolerate higher level of inconsistency in annotations and use ill-defined schemas. The system for semantic wiki authoring is presented in [2]. It aids users in specifying link types, while entering the wiki text. This approach considers ontology-like wiki types, using "is a" or "instance of" relationship types. Since the prototype supports manual editing, it does not discuss automatic relationship assignment. Our approach can be used as an additional feature in this system.

One of the first attempts to automatically extract the semantic information from Wikipedia is presented in [9], which aims at building an ontology from Wikipedia collection. This work focus on the extraction of categories using links and surrounding text, while we aim at extracting semantic links using assigned categories. The paper [7] shows the importance of automatic extraction of link types, and illustrates several basic link types, like synonyms, homonyms, etc. It also suggests to use properties for dates and locations. However, it does not propose any concrete solutions or experimental results. Studies of *history flow* in Wikipedia are presented in [11]. The work is focused on discovering collaboration patterns in page editing history. Using an original visualization tool they discovered editing patterns like statistical corroboration, negotiation, authorship, etc. This work does not consider semantic annotation of Wikipedia articles.

The link structures in Wikipedia have been studied recently. The work from [13] presents an analysis of Wikipedia snapshot on March 2005. It shows that Wikipedia links form a scale-free network and the distribution of indegree and outdegree of Wikipedia pages follow a power law. In [3] authors try to find the most authoritative pages in different domains like *Countries*, *Cities*, *People*, etc., using PageRank [5] and HITS [8] algorithms. It is reported in the paper, that Wikipedia forms a single connected graph without isolated components or outliers.

3 Problem

The usage of semantic links can be illustrated by the example we have mentioned in Section 1. Consider the query *find Countries which had Democratic Non-Violent Revolutions*. When we search in full-text for *Country Revolution Democracy* we get a lot of pages, which contain all the keywords, but most of them do not talk about particular countries. In a database-like view, the target page of our query should belong to the *Countries* category, and it should have a connection to a page in the category *Revolutions* which mentions the word *Democracy*. In current Wikipedia, there is actually a link between the pages *Ukraine* and *Orange Revolution*. If we put into a separate inverted list[1] all pages with *Country to Revolution* link type, we can force the previous query to return more relevant results.

However, it is infeasible to maintain and index all possible links between Wikipedia categories. An example of typical Wikipedia linkage between categories is shown in the Fig. 1. Ovals correspond to categories, squares contain the lists of pages and arrows show existence of at least on hyperlink between categories. The category *Republics* is pointed by the *Female Singers*, *Egg*, and *Non-violent Revolutions* categories. It also points to *Capitals in Europe*, *Spanish-American War People* and *Non-violent Revolutions* categories. Some of these links can be converted into strong semantic relationships, like "*Republics to Non-violent Revolutions*" categories, while relationships like "*Egg to Countries*" are not regular semantic connections and only used for navigation or some unimportant purposes. It is useless to type and index such "LinkSourceCatergory to LinkTargetCategory" relationships, as they cannot help users in search. Instead, we need to filter out unimportant links and extract semantically significant relationships from Wikipedia. This could be achieved by analyzing the link density and link structures between the categories.

Besides search, the prominent semantic relationships can be of use in template generation and data cleaning. For example, if we have some pages in *Countries* without link to pages in *Capitals*, the system could suggest users to add missing link.

One may want to create more precise link types and distinguish between type "Country has Capital" and "Country changed Capital". However, this task is much more chal-

[1] Inverted indices are used in information retrieval for keyword search, for detail see [14]

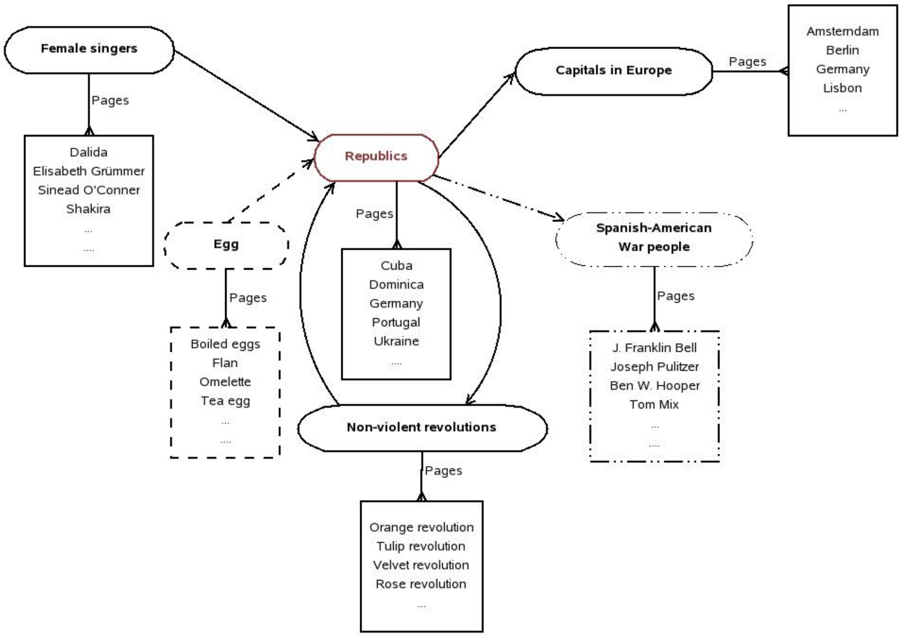

Fig. 1. The Wikipedia category Republics and several connected categories with corresponding sample pages. Arrows show the semantic connections between categories, dashdot lines show purely navigational links.

lenging and it is not the focus of this paper, in which we concentrate on selecting only coarse-grained semantic relationships.

4 Approach to Extracting Semantic Links

This section presents our approaches to extracting semantically important relationships from the links in Wikipedia. This task can be seen as an automatic construction of a database schema, where we want to emphasize the meaningful relationships between categories and disregard unimportant ones.

It seems reasonable, that highly connected categories represent strong semantic relations. For example, if a considerable percentage of pages from category "Country" have links to category "Capital", we can infer that there must be a "Country to Capital" relationship between the two instances categories. On the other hand, if there are only a few links between two categories like "Actor" and "Capital", it seems that there is no regular semantic relationship like "Actor to Capital".

We conduct experiments to test this filtering method. In the experiments, we extract a core set of pages which have a common topic (in our case the common topic is *Countries*). For these pages we extract all the categories they belong to, and also two lists of categories, one for the pages with links toward *Countries* (inlink pages) and one for the pages referred by *Countries* (outlink pages). The experiments with these lists can give an idea about what link direction is more important for semantic relationship discovery. During the experiments we test two measures used for finding the strong semantic connections:

1. **Number of links between categories**. The more links we have between pages in two categories, the stronger should their semantic connection be. As we study separately the effect of outgoing links and incoming links, each time only links in one direction are considered.
2. **Connectivity Ratio**. We can normalize the number of links with the category size, to reduce the skew toward large categories. We call this normalized value *Connectivity Ratio*, and it represents the density of linkage between two sets (in one direction). Namely
$$ConnectivityRatio_i = \frac{NL_{ij}}{NP_i}$$
where NL_{ij} is the number of links from category i to category j[2], and NP_i is the total number of pages in category$_i$.

We have received a valuable comment from anonymous reviewers, that size of the target directory is also important for normalization and NP_j could actually be included into the formula. We agree with this viewpoint and will experiment in future with more modifications of Connectivity Ratio.

5 Experimental Studies

In this section we describe our experiment setup and discuss the results.

5.1 Collection

For experiments we used the Wikipedia XML corpus [6] which is available for the participants of INEX 2006 evaluation forum[3]. This corpus is based on the English Wikipedia dump, it has about 668,670 pages, which belong to 63,879 distinct categories[4]; only pages from article namespace are included. We exported the dataset into a MySQL 4.1 database, the data size was about 1,2 Gigabytes.

For the experiments we selected three sets of pages, which we called *Countries*, *Inset* and *Outset*. The *Countries* set consists of 257 pages devoted to countries, they were manually extracted from the "List of countries" Wikipedia page, this set represents the *Countries* category. We did not use *Countries* category directly, since it contains subcategories like *European countries*, *African countries*, etc., rather than separate pages with countries. Since in this paper we do not consider hierarchical nature of categories,

[2] In current experiments j always corresponds to a *Countries* set.

[3] `http://inex.is.informatik.uni-duisburg.de/2006/`

[4] Some categories names differ only by space character before the names, or slightly different spelling. Our experimental setup does not assume use of NLP techniques, so we did not remove these inconsistencies and treated these categories as distinct.

we selected countries as described above. We also built the *Inset*, which contains all Wikipedia pages that point to any of the pages in the *Countries*, and *Outset* contains pages being pointed by the pages in *Countries*. The statistics summary for the selected sets is presented in Table 1.

	# of pages	# of assigned categories
Countries	257	405
Inset	289,035	60,277
Outset	30,921	14,587
Total (distinct entries)	290,893	63,879

Table 1. The Statistics from Experimental Collection

Each page consists of the name of the page, a list of associated categories, and a list of links that can be internal links (pointing to Wikipedia pages) or external links (pointing to pages from the Web). In our experiments, we only considered internal links.

5.2 Results

The main evaluation criteria for our task is the quality of extracted semantic relationships. To enable quantitative comparison between semantic connection, we introduce the **Semantic Connection Strength** measure (**SCS**). It receives value 0, 1 or 2, where value 2 represents a strong semantic relationship, value 1 represents a average relationship and value 0 represents a weak relationship [5]. In our assessment, the assessors were given the following instruction: "category A is strongly related to category B (value 2) if they believe that every page in A should conceptually have at least one semantic link to B; A and B are averagely related (value 1), if they believe 50% of pages in A should have semantic links to B; otherwise, A and B are weakly related (value 0)." This evaluation setup is slightly similar to one from [4], while we measure semantic connection between categories, rather than terms. Our experimental results showed that the level of disagreement between assessors could be high (sometimes it reached 40%). It indicates that SCS is a very subjective measure and should be improved in the future. In current experiments, only assessments made by one person were used, because we found important inconsistencies in other assessments and they could not be removed until the submission deadline.

In the first set of experiments, we tested whether the number of links between categories is a good indicator of the level of semantic relationship. By "number of links between categories" we mean the number of pages in source category, which have at least on link to any page in target category.

We ranked the categories from *Inset* and *Outset* by the number of pages in them, because according to the way we obtain *Inset* and *Outset*, it is exactly the ranking by number of links between categories. From each of the obtained rankings we selected 100 sample categories using a fixed interval, such that they are uniformly distributed

[5] The intermediate values are also possible when averaging the assessment results.

across each ranking. For example out of 15,000 we select categories number 1, 150, 300, 450, ..., 15000. These sample categories with corresponding numbers of links are listed in the Table 2.

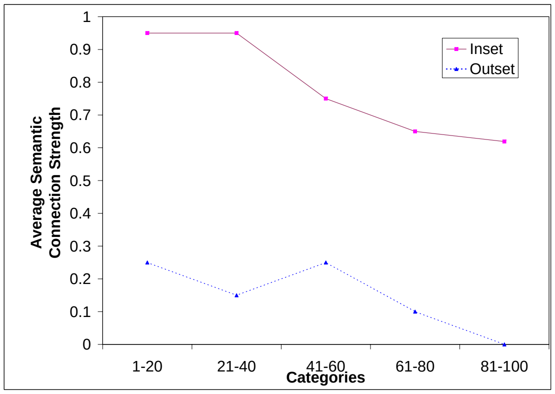

Fig. 2. Average semantic connections strength for 100 sample categories, extracted using number of links.

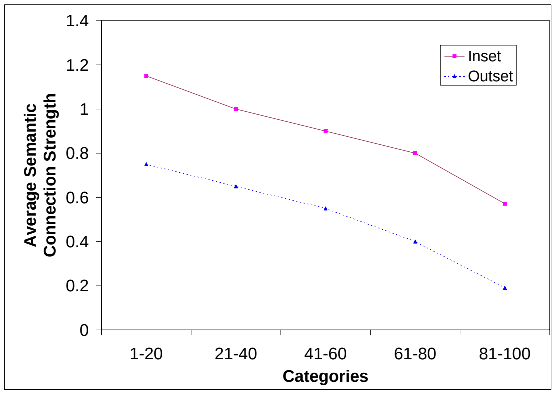

Fig. 3. Average semantic connections strength for 100 sample categories, extracted using Connectivity Ratio. A monotonic decrease shows a positive correlation between SCS and Connectivity Ratio.

The SCS measures of sample sets were averaged over every 20 categories, and the results are shown in the Fig. 2. On the ordinate we put the average of the SCS, and on the abscissa we show categories in descreasing order of number of links between pages in them and pages in Countries. We can see from the plot, that by using *Inset* we obtained stronger semantic relationships in comparison to *Outset*. This could be either

#	# of links	Inset	# of links	Outset
1	3272	American actors	193	Country code top-level domains
2	99	German poets	21	Governorates of Egypt
3	67	People from Arizona	15	South American history
4	52	1988 albums	12	Antigua and Barbuda
5	43	Rapists	10	Cote d'Ivoire
6	37	People from Hawaii	9	Yugoslavia
7	32	geography of Egypt	8	Ancient Japan
8	29	1974 films	7	Cross-Strait interactions
9	26	Stanford alumni	7	Empire of Japan
10	24	Camden	6	History of Mongolia
11	22	Pre-punk groups	6	Theology
12	20	Nuremberg Trials	5	Islands of Singapore
13	19	Video storage	5	Energy conversion
14	17	Neighbourhoods of Buenos Aires	5	Westminster
15	16	Dutch mathematicians	4	Geography of New Zealand
16	15	German currencies	4	Lists of lakes
17	14	National parks of Kenya	4	Yugoslav politicians
18	13	Cities in the United Arab Emirates	4	Encyclopedias
19	13	Egg	3	Subdivisions of Afghanistan
20	12	Swedish military commanders	3	Geography of Lebanon
21	11	Eurovision Young Dancers Competitions	3	Ecuadorian culture
22	11	Basketball at the Olympics	3	Rivers
23	10	Communes of Charente-Maritime	3	Nova Scotia
24	10	1846	3	Political parties in Sweden
25	9	New Zealand Reform Party	3	Roman Catholic Church
...	...	...	...	...
76	1	Australian sport shooters	1	Hindi
77	1	Canadian pathologists	1	Canadian television
78	1	Danish archbishops in Lund	1	Abstraction
79	1	Football in Uganda	1	Trinidad and Tobago writers
80	1	Ice hockey in China	1	Singaporean people
81	1	Latin_American_cuisine	1	Scythians
82	1	Mountains of Libya	1	Tetraonidae
83	1	Paradox games	1	Historic United States federal legislation
84	1	Road transport in Switzerland	1	Water ice
85	1	Spanish military trainer aircraft 1970-1979	1	Hiberno-Saxon manuscripts
86	1	Transportation in Manitoba	1	Belgian cyclists
87	1	Yom_Kippur_War	1	Business magazines
88	1	62 BC	1	British fantasy writers
89	1	Defunct Northern Ireland football clubs	1	Victims of Soviet repressions
90	1	Missouri Pacific Railroad	1	Empresses
91	1	U.S. generals	1	Medieval music
92	1	Creator deities	1	Soviet dissidents
93	1	Television stations in the Caribbean	1	University of Edinburgh alumni
94	1	Manitoba government departments and agencies	1	Signers of the U.S. Declaration of Independence
95	1	Libraries in Illinois	1	Microbiology
96	1	Star Wars Trade Federation characters	1	Communities in New Brunswick
97	1	Sin City	1	Electrical engineers
98	1	Ritchie County, West Virginia	1	Thomas the Tank Engine and Friends
99	1	Arapahoe County, Colorado	1	California Angels players
100	1	Mega Digimon	1	Towns in New Hampshire

Table 2. The 50 samples from category ranking built using number of links. The hypothesis is that categories should ordered by decreasing semantic strength of their connection to Countries.

#	Connectivity Ratio	Inset	Connectivity Ratio	Outset
1	1	Johannesburg suburbs	1	Provinces of Vietnam
2	1	Cities in Burkina Faso	1	New Zealand-Pacific relations
3	1	The Outlaws albums	1	Transportation in Lebanon
4	1	Gackt albums	1	9th century BC
5	1	North Carolina Sports Hall of Fame	1	Education in Belgium
6	1	Airlines of Liberia	1	Lake Kivu
7	1	Tongan rugby league players	1	Nepalese law
8	1	Hong Kong radio personalities	1	Sport in Lithuania
9	1	Yorb	0.928571	Republics
10	1	Education in Qatar	0.75	Economy of Greece
11	1	North African music	0.666667	Commonwealth Universities
12	1	Zara class cruisers	0.666667	World War II European theater
13	1	Airports in Shanghai	0.571429	Cities in Kosovo
14	1	Croatian athletes	0.5	Languages of Ukraine
15	1	Iranian photographers	0.5	Iraqi culture
16	1	Paleozoologists	0.5	1287
17	1	Swedish sportspeople	0.5	Bolivian music
18	1	Ceylon cricketers	0.5	Foreign relations of Hungary
19	1	Peanuts	0.5	Moroccan society
20	1	1997_films	0.5	Sri Lankan literature
21	1	Archaeological sites in Kazakhstan	0.461538	Politics of Macau
22	1	British make-up artists	0.416667	Ajaria
23	1	Coscoroba	0.4	Peninsulas of Russia
24	1	Farragut class destroyers	0.361111	Forced migration
25	1	High_schools_in_Florida	0.333333	Bessarabia
...	...	...	...	...
76	0.346154	BBC	0.0508475	Unitarian Universalists
77	0.333333	Hydrography	0.0487805	File sharing networks
78	0.333333	Porn stars	0.047619	Spanish Civil War
79	0.333333	Komsomol	0.0447761	Swedish nobility
80	0.32	1973 American League All-Stars	0.0425532	Battles of France
81	0.3	Roman Republic	0.04	Babylonia
82	0.285714	Dacian kings	0.0384615	Cantons of Switzerland
83	0.272727	University of San Francisco	0.037037	Scales
84	0.25	Esperantido	0.0344828	Agriculture organizations
85	0.25	Cooking school	0.0325203	Alcoholic beverages
86	0.242424	Church architecture	0.03125	New Testament books
87	0.222222	Danny Phantom	0.0294118	Marine propulsion
88	0.2	Buildings and structures in Cardiff	0.0273973	British politicians
89	0.2	Prediction	0.025641	Food colorings
90	0.181818	Media players	0.0238095	Christian philosophy
91	0.166667	Computer animation	0.0222222	Governors of Texas
92	0.153846	Kroger	0.0208333	West Indian bowlers
93	0.142857	Scottish (field) hockey players	0.0186916	Ancient Greek generals
94	0.125	Free FM stations	0.017094	Spanish-American War people
95	0.111111	Palm OS software	0.0155039	English ODI cricketers
96	0.09375	Stagecraft	0.0136986	Presidents of the Cambridge Union Society
97	0.0769231	Transportation in Texas	0.0117647	Municipalities of Liege
98	0.0625	Guessing games	0.00952381	Medical tests
99	0.0434783	Massachusetts sports	0.00714286	Food companies of the United States
100	0.0163934	data structures	0.00414938	Telecommunications

Table 3. The 50 samples from category ranking built using Connectivity Ratio. The hypothesis is that categories should ordered by decreasing semantic strength of their connection to Countries.

a sign of superior importance of inlinks or just show a special property of category Countries. We will try to answer this question in next set of experiments with more categories considered.

The better performance of *Inset* is also observed in the second set of experiments, where we used Connectivity Ratio as a ranking factor. The results are given in Table 3 and Fig. 3. The performance of the Connectivity Ratio measure is up to 25% better than that of number of links, which proves the advantage of the normalization.

The results are considerably less than 2, it shows that still a lot of weak semantic connections get to the top of the ranking and there is much space for improvement. The results are not round (0, 1 or 2) since they are averaged over intervals of 20 judgements.

We expected the Connectivity Ratio to rank semantically strong relationships higher and our pilot experiments supported this hypothesis. While current experiments are certainly not sufficient to prove general effectiveness of Connectivity Ratio on eny pair of categories, the monotonic decrease of both plots on Fig. 3 shows correlation between SCS and Connectivity Ratio, which means it worth working on it further. The important problem is to find relevant factors to include in category ranking algorithm, Connectivity Ratio behaves like a relevant factor for our ranking of categories by semantic relatedness.

6 Conclusions and Future Work

We have observed that, for a given category, inlinks have superior performance in comparison to outlinks. This could be either a sign of importance of inlinks or an evidence of a special property of category Countries. We will try to answer this question in next set of experiments with more categories considered.

We also show that normalized Connectivity Ratio is a better measure for extracting the semantic relationships between categories. We consider this result might be skewed toward our core *Countries* category, as it is natural that there are a lot of inlinks to the pages representing countries (consider that every event must happen in a country). The results we obtained are also influenced by the ranking scheme we chose. It is necessary to improve the Connectivity Ratio formula so that it can bring out more relevant relations and removes the trivial ones.

For our future experiments we want to select more categories as a starting set and remove bias introduced by the *Countries* categories. The assessment of semantic relationship should be improved by taking into account possible information need. It would be interesting to study a cardinality of link types relationships. For example, "Actor to BirthYear"[6] is a n:1 relation, while "Actor to Film" is a n:n relation. Another interesting aspect is to investigate bidirectional relationships, categories size and their indegree, we are also going to apply link analysis algorithms for establishing the semantic authorities among categories.

[6] In Wikipedia there are dozens of categories like "born 1970", "born 1971", etc., which represent persons who were born in particular year. We are grateful to anonymous reviewers, who made a good point that BirthYear is actually a relation, not the category. But we decided to keep this example to show common inconsistency of real-world data and to underline the difficulties one has to consider, while moving from theory to practice.

7 Acknowledgments

We would like to thank Paul Chirita for numerous discussions and Michal Kopycki and Przemyslaw Rys for their invaluable help with the experimental setup.

References

1. Wikipedia, the Free Encyclopedia. `http://wikipedia.org`, accessed in 2006.
2. David Aumueller. SHAWN: Structure Helps a Wiki Navigate. In *Proceedings of BTW Workshop WebDB Meets IR*, March 2005.
3. Francesco Bellomi and Roberto Bonato. Network Analisis for Wikipedia. In *Proceedings of Wikimania 2005, The First International Wikimedia Conference*. Wikimedia Foundation, 2005.
4. Abraham Bookstein, Vladimir Kulyukin, Timo Raita, and John Nicholson. Adapting Measures of Clumping Strength to Assess Term-Term Similarity. *Journal of the American Society for Information Science and Technology*, 54(7):611–620, 2003.
5. Sergey Brin and Lawrence Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
6. Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML Corpus. Technical report, 2006.
7. Daniel Kinzler. WikiSense — Mining the Wiki. In *Proceedings of Wikimania 2005, The First International Wikimedia Conference*. Wikimedia Foundation, 2005.
8. Jon Kleinberg. Authoritative sources in a hyperlinked environment. Technical Report RJ 10076, IBM, 1997.
9. Natalia Kozlova. Automatic Ontology Extraction for Document Classification. Master's thesis, Saarland University, Germany, February 2005.
10. Markus Krötzsch, Denny Vrandecic, and Max Völkel. Wikipedia and the Semantic Web - The Missing Links. In *Proceedings of Wikimania 2005, The First International Wikimedia Conference*. Wikimedia Foundation, 2005.
11. Fernanda B. Viegas, Martin Wattenberg, and Kushal Dave. Studying Cooperation and Conflict between Authors with History Flow Visualizations. In *Proceedings of SIGCHI 2004, Vienna, Austria*, pages 575–582. ACM Press, 2004.
12. Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic Wikipedia. In *Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland*, 2006.
13. Jakob Voss. Measuring Wikipedia. In *10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, 2005.
14. Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, 1999.