An Approach for Deriving Semantically Related Category Hierarchies from Wikipedia Category Graphs

Khaled A. Hejazy and Samhaa R. El-Beltagy

Center for Informatics Science, Nile University, Cairo, Egypt
khaledhejazy86@yahoo.com, samhaa@computer.org

Abstract. Wikipedia is the largest online encyclopedia known to date. Its rich content and semi-structured nature has made it into a very valuable research tool used for classification, information extraction, and semantic annotation, among others. Many applications can benefit from the presence of a topic hierarchy in Wikipedia. However, what Wikipedia currently offers is a category graph built through hierarchical category links the semantics of which are undefined. Because of this lack of semantics, a sub-category in Wikipedia does not necessarily comply with the concept of a sub-category in a hierarchy. Instead, all it signifies is that there is some sort of relationship between the parent category and its sub-category. As a result, traversing the category links of any given category can often result in surprising results. For example, following the category of "Computing" down its sub-category links, the totally unrelated category of "Theology" appears. In this paper, we introduce a novel algorithm that through measuring the semantic relatedness between any given Wikipedia category and nodes in its sub-graph is capable of extracting a category hierarchy containing only nodes that are relevant to the parent category. The algorithm has been evaluated by comparing its output with a gold standard data set. The experimental setup and results are presented.

Keywords: Wikipedia, Semantic relatedness, Semantic similarity, Graph analysis, Category hierarchy, Hierarchy extraction.

1 Introduction

Wikipedia is an online encyclopedia that has more than 23,000,000 articles in which, more than 4 Millions articles are in English covering a wide variety of topics. Articles are maintained by more than 100,000 active volunteer contributors. As Wikipedia is written collaboratively by anonymous volunteers, anyone can write and change Wikipedia articles. It is assumed that contributors will follow a set of policies and guidelines developed by the Wikipedia community. However, there is nothing in place to enforce editing policies before or during contributing[1] which means that breaches to Wikipedia's policies and guidelines are being conducted by its community, greatly affecting its quality.

[1] http://en.wikipedia.org/wiki/Wikipedia

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 77–86.
DOI: 10.1007/978-3-642-36981-0_8

Just like articles, Wikipedia's categories are socially annotated. When creating new categories and relating them to previously created ones, there is no strict enforcement of which higher-level categories a child sub-category can belong to; thus, Wikipedia's category structure is not a tree, but a graph in which links between nodes, have loosely defined semantics.

Consequently a sub-category in Wikipedia does not necessarily comply with the concept of a sub-category in a hierarchy. A category label in Wikipedia is simply intended as a way for users to navigate among articles, and only signifies that there is some sort of a relationship between the parent category and its sub-category that is not necessarily of the type "is-a" which is expected in a hierarchical Knowledge Organization System. This problem causes irregularity in semantics between categories that is amplified in deeper levels. For example, following the category of "Computing" down its sub-category links, the totally unrelated category of "Theology" appears. Also, the graph nature of the Wikipedia category structure means that following the sub-category links of any given category, can eventually lead back to the same category. Detecting and eliminating cycles is a minor issue. Detecting sub-categories that should be considered as belonging to any given category is the main challenge addressed by this work. To address this challenge, an approach for measuring lexical semantic relatedness between Wikipedia's categories and nodes in their sub-graphs and using this as an indicator for relatedness, was developed.

In this paper, we introduce this new approach for deriving semantically related category hierarchies from Wikipedia category graphs and extracting a category hierarchy containing only sub-categories that are relevant to the parent category.

The rest of the paper is organized as follows; firstly, related work is presented in section (2), the proposed approach is described in section (3), the procedure followed for evaluating our approach and the experimental results are presented in section (4). Analysis of the results is discussed in section (5). And finally section (6) concludes this paper.

2 Related Work

Since its inception, Wikipedia has undergone tremendous growth, and today it is the largest online encyclopedia known to date. Wikipedia has been widely used as a huge resource of concepts and relationships for text mining tasks; like classification, information extraction, and computing semantic relatedness of natural language texts, among others. Most research works that make use of Wikipedia have used Wikipedia's concepts and relationships as is, except for some preprocessing and slight modifications. No previous research (as far as the authors are aware) addressed semantic irregularity between categories in Wikipedia's categorization system.

Wikipedia's categories' growth has previously been analyzed in [1], where an algorithm that semantically maps articles by calculating an aggregate topic distribution through the articles' category links to the 11 top Wikipedia categories (manually selected). Semantic relatedness for category nodes is then calculated through link distance metrics, such as the length of the shortest path between two nodes.

The evolution of Wikipedia's category structure over time has been studied in [2]. Results of this research have shown that the Wikipedia category structure is relatively stable for a bottom-up evolved system. However, the work did not address the accuracy of the category structure.

Wikipedia has been used for measuring lexical semantic relatedness between words or text passages. Explicit Semantic Relatedness (ESA) [3] has been shown as a successful measure for semantic relatedness. It treats each Wikipedia article as a dimension in a vector space. Texts are then compared by projecting them into the Wikipedia articles' space, then measuring the similarity between vectors using conventional metrics like cosine similarity. Because this work relies mostly on individual articles, the category structure of Wikipedia was not an issue.

Wikipedia has been used to compute semantic relatedness by taking the categorization system of Wikipedia as a semantic network [4].

Wikipedia Link-based Measure [5] also measures the semantic similarity of two Wikipedia pages by comparing their incoming and outgoing links. The score is determined using several weighting strategies applied to the overlap score of the articles' links.

In this paper, we propose an approach for deriving semantically related category hierarchies from Wikipedia category graphs. Our approach is somehow similar to ESA, except the fact that we are measuring semantic relatedness between categories instead of articles or words. Also, we use a key-phrase extraction for dimensionality reduction.

3 Methodology

Detecting semantically related categories based on measuring lexical semantic relatedness between them requires an efficient representation for each category. A TF-IDF scheme [6] has been used to assign weights to the feature vectors representing Wikipedia categories. In the following subsections, we start with the pre-processing step; in which we discuss the data sources with their components and the pre-processing steps conducted before these data are used, and then we discuss the steps of generating the feature vectors of Wikipedia categories.

3.1 Pre-processing

Wikipedia's backups are created regularly by the Wikimedia Foundation[2]. These dumps are publicly available. We have used Wikipedia's XML dump release 02-05-2012, which contains all Wikipedia article pages. The size of the uncompressed dump is around 38 GB.

Pages in this xml dump are represented by multiple tags. From those our system uses the page's unique ID, page's title, page's time stamp, and page's text.

[2] www.wikimedia.org

In order to handle this large XML dump file, apache Solr [7] has been used to index it through its Data Import Handler, which also facilitated the searching processes required by our followed approach.

We also acquired some relevant SQL files from the same source in order to allow us to re-construct the categorization graph of Wikipedia. One of these files is the enwiki-20120502-category.sql.gz which is an SQL file containing metadata for each category in Wikipedia; its category ID (differs from the page ID), its title, and the number of its pages and subcategories. The other is 20120502-categorylinks.sql.gz which is an SQL file has been acquired, and used for building Wikipedia's categorization graph. The SQL file contains the page IDs of any page defined as a category member, the page title of the category's description page, the time stamp of the approximate addition time of the link, the category link type that determines whether the page ID is a page, a sub-category or a file, along with some other attributes for sorting and for defining the collation of the category links.

Pages in Wikipedia are not only articles; categories are special pages that are used to group articles together, and their titles start with the namespace "Category:". Also, there are administrative pages with different namespaces that are not used to share encyclopedic information, but rather to preserve policies created and implemented

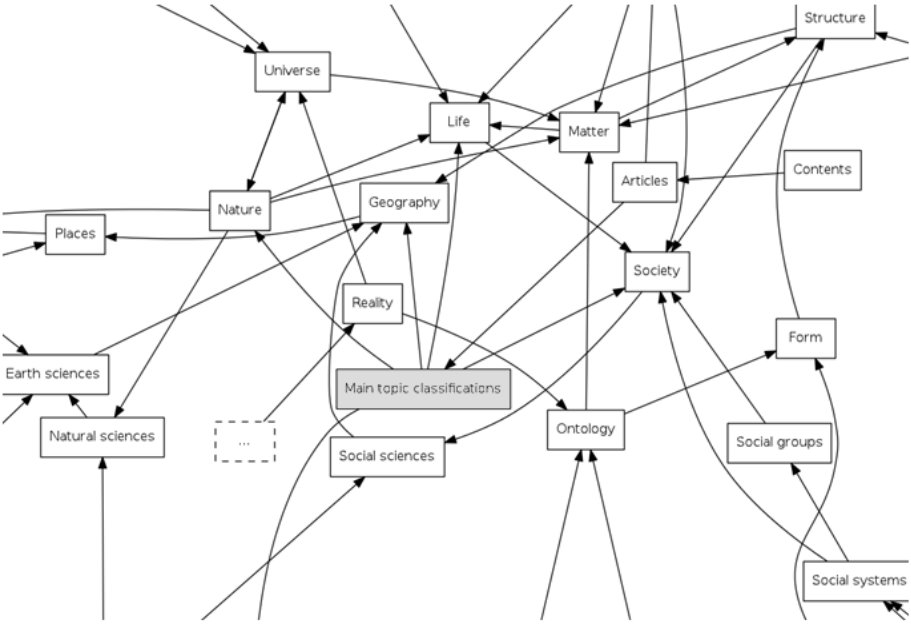

Fig. 1. A sub-graph in Wikipedia showing the category "Main topic classifications" and some of its super-categories and sub-categories

by user consensus for editing Wikipedia, such as User, User talk, File, File talk, Template, Template talk, Help, Help talk, Portal, Portal talk, etc.[3]. A cleanup step has been conducted for removing all administrative pages from the XML dump, leaving behind only article and category pages.

The Wikipedia category system does not offer top level categories. So, the category "Main topic classifications" was selected as our root category, the sub-categories of which are topical categories, indirectly contain almost all Wikipedia articles. Figure 1 shows an example of the category "Main topic classifications" and some of its parent and children categories. The parents, as well as the sub-categories, have many other children that are not shown in the figure.

3.2 Algorithm

The approach that is followed in this work is based on the observation that each domain or category has its own unique vocabulary. For example, the terms {north, region, island, earth surface, river, geography, south, geographical, urban, landscape, map, spatial, earth, sea, etc} collectively represent a subset of terms that are often used in the context of geography. Sub-categories that are in fact the hierarchical descendants of any given category are highly likely to share the same vocabulary and use the same terminology. Direct sub-categories are more likely to be related to their parent categories than their 2nd level or other deeper level descendants.

Basically, the relationship between a root category and its descendents grows weaker as we go deeper down the tree. Having said that, even first level sub-categories in Wikipedia can sometimes be un-related to their parent category in the hierarchical sense. The assumption made by this work, is the categories that are directly related to some parent category, will most likely share a reasonable part of its vocabulary. A category that has an entirely different vocabulary is not likely to be a hierarchical decedent of its parent category, even if some other relationship exists between the two. To build on this observation, two steps had to be followed:

1. Modeling the vocabulary for a given category
2. Measuring semantic relatedness between categories

Modeling the Vocabulary of a Category

In order to make use of a category's vocabulary, the first step that needed to be carried out was to capture the vocabulary of Wikipedia categories. In Wikipedia, each category has both a set of direct pages and a set of direct descendents or sub-categories. The set of pages is sometimes very small, or non-existent, making it very difficult to model the vocabulary of the category based entirely on these. First level categories, while often related to a category, also often have some noise. In the proposed approach, both a category's pages and its immediate sub-categories count towards building the category's vocabulary by extracting key-phrases from both. In the context of this work, key-phrases are defined as a list of terms each of which is made up of one or more words and that describe the "sub-category" with which they are associated.

[3] http://en.wikipedia.org/wiki/Wikipedia:Administration

The KP-Miner system [8] was the system used to extract key-phrases. When applied to a group of Agricultural documents, 66% of the key-phrases extracted from those, were found to correspond to concepts in the Agricultural Ontology known as AGROVOC4 and 90% were found to be directly related to the field of Agriculture [8]. The system which was designed to be generic relies on term frequency information gathered from a large corpus of random pages. To make it more relevant to Wikipedia, the system was re-trained using Wikipedia's articles, in the sense that term frequency information was obtained based on term occurrences within Wikipedia. The system was then used to extract the top n-key-phrases from a given text in the manner described below.

In order to model the vocabulary of a category, a preliminary step of extracting n key-phrases from each of Wikipedia categories' direct pages (if possible) is conducted. To extract the key-phrases, all pages are concatenated and treated as a single document; extracting n key-phrases referred to as "*Pages_Keyphrases*", which has been indexed and stored in a multi-valued text field in Solr. The number of key-phrases extracted to from the *Pages_Keyphrases* of each of Wikipedia categories is fixed as a constant $\boldsymbol{n}$ ($\boldsymbol{n}$ = 300), regardless of the size of a category, to prevent larger categories from biasing the model of their parent category (in the future, we intend to experiment with different values of $\boldsymbol{n}$).

In order to build the representative feature vector for each of Wikipedia's category (referred to as "*Category_Keyphrases*"); direct pages of each category are considered as the most important resource for representing the category; that's why *Pages_Keyphrases* are fully included in the *Category_Keyphrases*. Also the direct sub-categories' *Pages_Keyphrases* were used for constructing a Category's *Category_Keyphrases* vector in a way that amplifies the common concepts among sub-categories, and excludes noise that can appear in any of them.

Each key-phrase obtained from each sub-category's *Pages_Keyphrases* can actually be thought of as a single vote for this key-phrase. Only key-phrases with votes greater than some value $\boldsymbol{m}$ (obtained from all 1st level subcategories), are included in the *Category_Keyphrases* vector of the parent category along with those of its *Pages_Keyphrases*, which then serves as a representative for that category.

Table 1 and Table 2 show the extracted *Pages_Keyphrases*, and *Category_Keyphrases* for the category "Islands". To calculate the weight of each key-phrase, both its frequency and its IDF factor are used.

Table 1. Sample of the key-phrases obtained for the *Pages_Keyphrases* of the category "Islands"

Stemmed Key-phrases		
island	unsinkable aircraft carrier	islet
floate island	island ecosystem	new zealand
coral reef	island restoration	reef
artificial island	private island	high island
unsinkable aircraft	low island	oceanic island

Table 2. Sample of key-phrases of the *Category_Keyphrases* for the category "Islands"

Stemmed Key-phrases		
island	isle	south
sea	archipelago	pacific
area	pacific ocean	coral reef
population	reef	indian ocean
map	ocean	sea level

In classical information retrieval models, the frequency of a term is calculated as the number of times it appears in a document. This is often normalized by dividing that number by the total number of terms that appear in the same document. In our proposed approach, a category is treated as a single document. The frequency of a key-phrase is calculated as the total number of times that this key-phrase has occurred across its sub-categories and its pages, and the weight is determined by multiplying this value with the IDF value obtained across all obtained key-phrases from Wikipedia.

Measuring Semantic Relatedness between Categories

After obtaining all feature vectors for all Wikipedia categories, building a hierarchical tree for any category becomes possible. To build such a tree for any category, its subcategories are traversed in a depth first fashion in order to accept or reject as hierarchical descendents of the category in question.

A subcategory is said to be accepted if the cosine similarity of its vector and that of the main category under consideration is greater than an empirically calculated threshold Ω.

4 Evaluation

4.1 Building the Evaluation Dataset

Humans have the natural ability to disambiguate topics and judge their relatedness. In order to evaluate our algorithm, a test dataset of 1000 categories has been randomly collected from the sub-graph of the category "Geography" in Wikipedia. Each instance represents a Wikipedia category that may or may not be considered as a semantically related sub-category to the main category being tested ("Geography" in our case). The test dataset was then manually annotated by 3 different human judges; determining whether or not semantic relatedness exists between each of the testing sub-categories and the main category being tested. The final manual annotation for each instance was determined by taking the consensus annotation represented by having the majority votes of the 3 judges. The resulting dataset[4] was used as a gold standard.

[4] The dataset is available upon request, and it will be available shortly on our website `http://tmrg.nileu.edu.eg/`

4.2 Results

The developed system was used to derive the hierarchical tree of the Geography category and the results were compared with the gold standard dataset described in the previous section. Table 3 shows the different results of the algorithm when compared against the gold standard dataset using different values for the semantic relatedness threshold Ω.

Table 3. Evaluation of the algorithm with different thresholds for Ω

#	Ω	Precision	Recall	F-Score
1	0.076	0.51078167115903	0.844097995545657	0.636439966414778
2	0.086	0.519662921348315	0.824053452115813	0.637381567614126
3	0.096	0.535871156661786	0.815144766146993	0.646643109540636
4	0.11	0.544753086419753	0.78619153674833	0.643573381950775
5	0.12	0.547049441786284	0.763919821826281	0.637546468401487
6	0.13	0.552238805970149	0.741648106904232	0.633079847908745
7	0.14	0.573476702508961	0.712694877505568	0.635551142005958
8	0.146	0.5893536121673	0.690423162583519	0.635897435897436

It was found that setting the threshold Ω to be 0.096 gives the highest F-score value. Thus, the following analysis section focuses on analyzing the results while setting Ω to be 0.096.

5 Analysis

As shown in table 3, there is a tradeoff between the precision and the recall; increasing the threshold results in increased precision and decreased recall, and vice versa.

Tables 4, and 5, show a sample of discrepancies between results of the presented system and manually annotated data, while tables 6, and 7 show samples of agreement between the two. The term "ACCEPTED" is used for indicating that the developed system has concluded that the sub-category in question is semantically related to the category being tested ("Geography" in this case) and that it should be part of its subtree, while the term "REJECTED" is used when there is no semantic relatedness.

Looking at table 4, and taking the category "Mountaineering" [5] as an example, it is easy to see why this category was mistakenly accepted. Mountaineering is a sport, or a hobby of mountain climbing, however, there is an overlap between the "Mountaineering" concept and the geographical concepts like "climbing mount Everest that is located in somewhere between China and Nepal". Taking another example of category "Paços de Ferreira" [6] from table 5, to examine why this category was rejected, when it should have been accepted, we find that this particular category does not have any sub-categories and only 3 pages the textual content of which is too poor to extract meaningful and sufficient key-phrases from.

[5] http://en.wikipedia.org/wiki/Category:Mountaineering
[6] http://en.wikipedia.org/wiki/Category:Paços_de_Ferreira

Table 4. Examples of categories detected as ACCEPTED while they are manually annotated as REJECTED

Category	Depth	Cosine Similarity
Mountaineering	5	0.2246
Women who reached the Poles	4	0.177
Lists of buildings and structures	3	0.154
Telecommunications infrastructure	4	0.17956
Baltimore City College	5	0.099

Table 5. Examples of categories detected as REJECTED while they are manually annotated as ACCEPTED

Category	Depth	Cosine Similarity
Underground cities	4	0.056
Ramsar sites in Israel	5	0.07657
Protected areas of the Republic of the Congo	5	0.0477
Paços de Ferreira	5	0.0154
Kronoberg County	5	0.0604

Table 6. Examples of categories detected as ACCEPTED and manually annotated as ACCEPTED

Category	Depth	Cosine Similarity
Geography of Austria	5	0.16888
Pas-de-Calais	5	0.156
Barnsley	5	0.14856
Brighton and Hove	5	0.1453
Algarve	5	0.15865

Table 7. Examples of categories detected as REJECTED and manually annotated as REJECTED

Category	Depth	Cosine Similarity
The Chronicles of Narnia music	5	0.0372
Yorkville University	5	0.0339
People from the Azores	5	0.0716
Science and technology in Uganda	5	0.04715
Health in Cyprus	5	0.01122

6 Conclusion

This paper presented a novel approach for deriving semantically related category hierarchies from Wikipedia category graphs. Future work will focus on refining the developed methodology so as to improve both precision and recall.

This approach is being applied within an ongoing project to generate a semantically related category hierarchy for collecting statistics on Wikipedia categories (where a category refers to an entire hierarchy) based on their number of pages, language instances, in-links, and out-links, among others. The statistics generated based on this hierarchy are supposedly more real than those generated from Wikipedia's category system.

Acknowledgements. This work was partially funded by Microsoft's Advanced Technology Lab in Cairo, grant number CIS-001-1011. Special thanks to members of NU's Text Mining research group for their insightful comments and help.

References

[1] Kittur, A., Chi, E.H., Suh, B.: What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1509–1512 (2009)
[2] Suchecki, K., Salah, A.A.A., Gao, C., Scharnhorst, A.: Evolution of Wikipedia's Category Structure. Advances in Complex Systems 15 (2012)
[3] Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1606–1611 (2007)
[4] Strube, M., Ponzetto, S.P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence, vol. 2, pp. 1419–1424 (2006)
[5] Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. Association for the Advancement of Artificial Intelligence (2008)
[6] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
[7] Apache Solr (2012), `http://lucene.apache.org/solr/`
[8] El-Beltagy, S.R., Rafea, A.: KP-Miner: A keyphrase extraction system for English and Arabic documents. Information Systems 34(1), 132–144 (2009)