Wikipedia World Map: Method and Application of Map-like Wiki Visualization

Cheong-Iao Pang
Dept. of Computer and Information Science
Faculty of Science and Technology
University of Macau
Macau S.A.R., China
ma76543@umac.mo

Robert P. Biuk-Aghai
Dept. of Computer and Information Science
Faculty of Science and Technology
University of Macau
Macau S.A.R., China
robertb@umac.mo

ABSTRACT

Wiki are popular platforms for collaborative editing. In volunteer-driven wikis such as Wikipedia, which attracts millions of authors editing articles on a diverse range of topics, contributors' editing activity results in certain semantic coverage of topic areas. Obtaining an understanding of a given wiki's semantic coverage is not easy. To solve this problem, we have devised a method for visualizing a wiki in a way similar to a geographic map. We have applied our method to Wikipedia, and generated visualizations for several Wikipedia language editions. This paper presents our wiki visualization method and its application.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*web-based services*; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*collaborative computing*; I.3.8 [**Computer Graphics**]: Applications

General Terms

Design, Experimentation

Keywords

Wikipedia, information visualization, category, semantic coverage

1. INTRODUCTION

In recent years wikis have become popular platforms for collaborative editing in many application domains, from managing projects to facilitating knowledge management and others. The most well-known public wiki site is the free online encyclopædia Wikipedia [13]. Wikis usually share the same characteristics: allowing people to publish content, and providing mechanisms for classifying content into topic areas, i.e. categories. These operations are performed manually by the wiki's community of users. As a result a wiki is the product of large-scale human collaboration.

Wikipedia has great value that has not yet been fully researched. Past research on Wikipedia has focused on both *micro-level* and *macro-level* of analysis. A micro-level of analysis typically focuses on a single article, whereas a macro-level of analysis studies the wiki as a whole, exploring relationships and the evolution of the entire content collection, among others. Our research falls in the latter class. In this project we aim to obtain an overview of Wikipedia and identify popular topic areas. By applying this to different language Wikipedias we wish to discover differences among those different language editions, and by implication differences of interest in those topic areas among the user communities of those language groups. However, our aim in this research is for our methods and tools to be general enough to be applied to other wikis besides Wikipedia, so that for example they could be used on an intra-organizational wiki as well.

Obtaining the above information from a wiki is a challenging task. Taking the English Wikipedia as an example, it is entirely impractical to analyse the raw data manually since it currently (August 2011) involves over 3.7 million articles and a terabyte of data in the database[1]. Although some data analysis and visualization tools for wikis exist, the output produced by many of these tools is usually difficult to understand by untrained users. The goal of our research thus is to create a method for creating an overview visualization for wikis that can produce a quasi "world map" with an appearance similar to a geographic map (but with no correspondence of this "wiki world" to our physical world), as even untrained end-users can usually readily understand and relate to such maps. We map elements of the wiki to visual elements of a traditional geographic map. For instance, wiki categories are represented as "countries", sub-categories as "regions", and articles as "cities". Besides end-users, wiki researchers can also benefit from such a visualization by obtaining an easily understandable overview of a wiki for analysis.

The remainder of this paper is organized as follows: Section 2 briefly presents related work. Section 3 discusses the pre-processing of wiki data, and Section 4 the visualization method itself. In Section 5 we discuss applications of the visualization, and make conclusions in Section 6.

2. RELATED WORK

Wikis, as well as Wikipedia, are growing both in size and value. Thus they attract focus from numerous researchers in different fields worldwide. On the other hand, research in information visualization is also growing rapidly. This section gives a brief review of pertinent research in both areas.

2.1 Wiki Category Pre-processing

Wikipedia, the wiki system that is the focus of this paper, has a category organization that is similar to a tree structure. However,

[1] http://meta.wikimedia.org/wiki/List_of_Wikipedias

WikiSym'11, October 3-5, 2011, Mountain View, CA, USA.

because the creation and maintenance of the category structure is a manual task performed by Wikipedia users a small number of cases such as multiple parents, loops, and other anomalies exist. Indeed the category structure of Wikipedia can be classified as a kind of *directed acyclic graph* rather than a tree. However, trees are more preferable to use in many cases for their simplicity, therefore several studies have developed methods to transform the category structure of Wikipedia into a tree.

To solve the cases of multiple parents, Yu et al. remove multiple repeated parents in sub-categories using Dijkstra's shortest path algorithm, by keeping the parent which is closest to the root and discarding the other. Whenever multiple parents are found, a TF-IDF cosine similarity measurement is applied to candidate nodes, in order to select the one that is most relevant to the child [18]. TF-IDF cosine similarity is a method to determine relevance of two documents by using the frequency of words occurring in both.

Zesch and Gurevych suggest a simple mechanism to solve the problem of loops. They process the categories in Wikipedia as a graph, and then use a depth-first search to traverse the category graph. Whenever cycles are detected among the nodes of the same level, they simply remove one of the links on that level to eliminate the cycles [19].

2.2 Category Similarity Calculation

Our visualization method creates an overview of a wiki mainly based on the relationship among categories. Once we know this relationship we can determine the individual positions of all categories in the drawing plane. The relationship of categories can be visually represented by their proximity, i.e. similar categories are placed close to each other.

Holloway et al. introduce a method for computing similarities among wiki categories by using the number of *co-assignments* of the same categories in articles [5]. Assuming that an article is assigned with the categories related to its content, an article acts as a connection between a pair of categories. In this way, a larger number of this connection (i.e. co-assignment) implies a stronger relationship between categories. Cosine similarity has been used for a long time in computing similarity between articles linked by identical keywords [1, 10, 14], but it is innovative to apply the method for calculating category relationships.

2.3 Wiki Visualization

Information visualization helps people understand complicated and abstract data, especially for large amounts of data such as in Wikipedia. Therefore increasing numbers of researchers have developed methods to visualize a wiki. Some of them focus on a single article, for example history flow visualization, visualizing the evolution of different revisions of an article [16]. Another type of visualization aims at giving an overview of an entire wiki or a part of it, such as category visualization. Holloway et al. render wiki categories as dots of different colours, representing the semantic coverage which is formed by categories. Figure 1 shows an example of this visualization [5] (dots represent categories, dots in colours other than grey represent selected categories as indicated in the legend). Some types of visualization focus on analyzing users' activities and authorship. Wattenberg et al. created an application called Chromograms [17] which displays operations performed on the content of Wikipedia, such as spell-checking, writing new content, reverting changes, etc.

2.4 Map-like Visualization

Most people understand geographic maps easily. Elements such as mountains, valleys, land, sea, rivers, and cities, as well as the meaning of each, are readily recognized by people even without special training. Therefore visualizing information structures in the form of a geographic map enables people to relate to such representations more easily without requiring prior instruction.

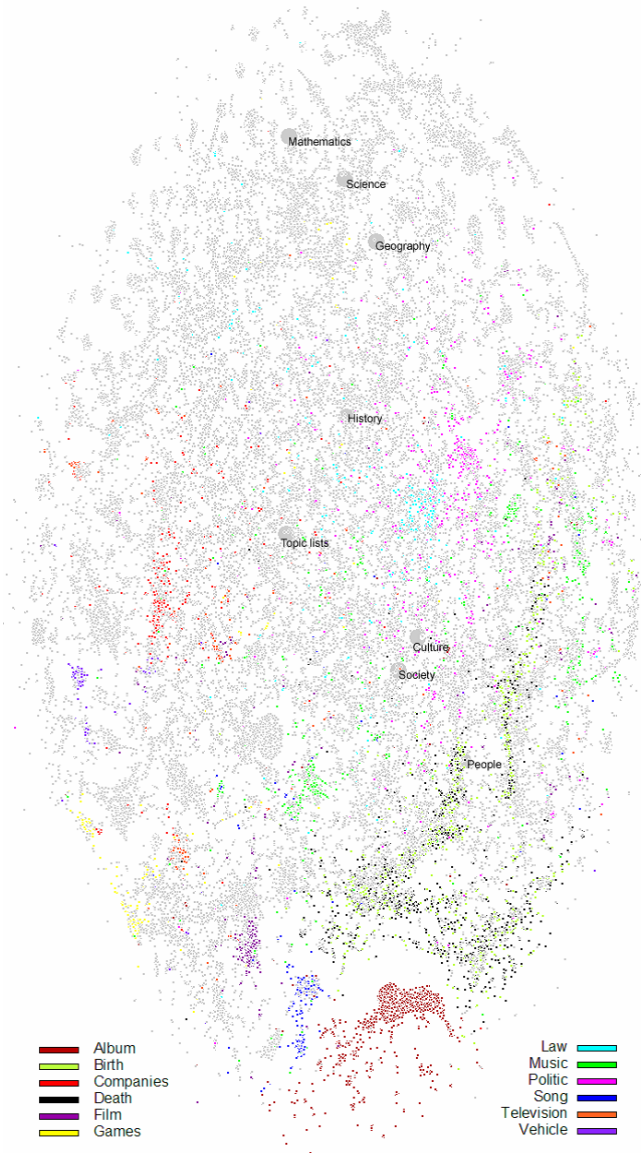
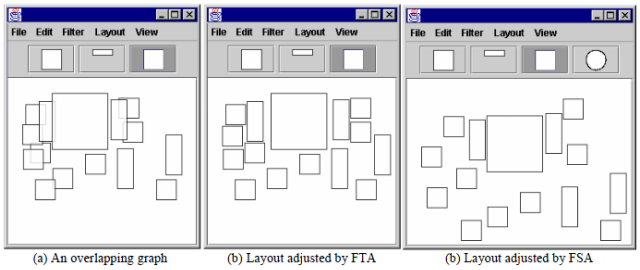


Figure 1: Wikipedia category visualization (reproduced from Holloway et al. [5])

Skupin presents a method that produces a map-like visualization for a knowledge domain based on the *Self-Organizing Map (SOM)* algorithm. SOM is a type of artificial neural network, where data is fed in and organized through an unsupervised learning process. The outcome of SOM is a low-dimensional map that represents the multi-dimensional input data [9]. Skupin's method was novel to apply the SOM algorithm to create a map-like visualization. Data is first transformed into a set of vectors in multiple dimensions, and then vectors are fed into the SOM to obtain a preliminary result. The preliminary result is then filled into a lattice of hexagons, followed by adding borders and text labels to finalize the visualization (Figure 2, which shows topic areas in the form of a geographic map).

Figure 2: Visualization of a 10-year period by cartographic means (reproduced from Skupin [15])



(a) An overlapping graph (b) Layout adjusted by FTA (b) Layout adjusted by FSA

Figure 3: Example of overlap removal algorithms (reproduced from Huang et al. [6])

2.5 Overlap Removal Algorithms

Overlap removal algorithms play an important role in map visualizations. Firstly, a generated layout of areas may contain overlapping areas. Moreover, text labels or other elements in a map may also overlap one another. As a result overlap removal algorithms are applied to reduce or eliminate such problems.

An early algorithm for overlap removal is the *Force-Scan Algorithm (FSA)* [12]. Overlaps cannot be removed without moving one of the overlapped items. The algorithm maintains relative positions of items compared to those before movement, but distances between items may be changed. The force-scan algorithm performs its task by scanning overlapped items in both horizontal and vertical directions, finding values of forces which are able to push/pull items apart in order to eliminate overlaps.

Huang et al. suggested a new algorithm called *Force-Transfer Algorithm (FTA)* for overlap removal [6]. The new algorithm adds an enhancement to FSA in compacting the space between items in the output, while still maintaining the same effectiveness as FSA. Figure 3 shows the comparison between adjustments of overlaps with both algorithms.

2.6 Research on Wikipedia Analysis

In recent years Wikipedia has been serving as a laboratory for researching its content and the social collaborative authoring process. Holloway et al. created a visualization to analyse the topic distribution in Wikipedia [5]. Kittur analysed conflicts of editing, in addition to also studying the topical classification in Wikipedia [8]. These researches give an overview on Wikipedia's content which are a foundation for other research such as our own.

Another area of research looks into the relationship between different language editions of Wikipedia. Since articles with the same concepts but written in different languages can be linked across different language Wikipedias, thus understanding such linkages can help draw conclusions on editors' behaviour in different parts of the world and even derive patterns of their different focuses on article topics. Hecht et al. examined diversity of knowledge representations and users' bias in recent works [2, 3].

From the revision history and related user information such as IP addresses, researchers are able to obtain users' location, as demonstrated by Lieberman [11]. Based on this work, Hecht shows the degree of relationship between user-contributed content and their location [4].

Table 1: Sizes of Wikipedia language editions processed

Language	Number of Articles
Danish	145,000+
Chinese	348,000+
Swedish	380,000+
German	1,055,000+
English	3,411,000+

3. DATA PRE-PROCESSING

This and the following section discuss our method for creating a map-like visualization. Before generating the visual outcome, data pre-processing is the first step to be performed.

We apply our method to Wikipedia which is based on the MediaWiki system, and have tested it with several Wikipedia language editions. However, concepts and steps of the method should be generic enough to be applicable on other wiki systems with only minor or no modification.

3.1 Wikipedia Data Source

Our data source are the publicly available database dumps provided by the Wikimedia Foundation[2]. Our method only requires two of the MediaWiki database tables, namely "categorylinks" and "page", rather than a full database dump. This greatly reduces our storage and processing requirements.

We processed and compared data from database dumps of five different language editions: Danish, Swedish, Chinese, German and English, all from January 2011. The approximate numbers of articles in the dumps of these language editions are shown in Table 1. We selected these database dumps for several reasons: firstly because we are conversant in these languages (which is needed for interpreting the visualized result); but also to give us a selection of very large (English), medium-sized (Chinese, Swedish) and small (Danish) Wikipedias; moreover, the English version is about three times the size of the German one, which in turn is about three times the size of both the Chinese and the Swedish versions. This enables us to compare wikis of a different scale as well as pairs of wikis of a similar relative size difference.

3.2 Category Graph Transformation

The data structure of categories in Wikipedia is a *directed acyclic graph*, starting from a root node. Sub-categories are created manually and connected to the parent node afterwards. Every category node can connect to any other nodes without restriction, thus potentially leading to problems such as multiple parents and loops.

[2]http://dumps.wikimedia.org/

Table 2: Top content categories in different Wikipedia language editions

Swedish	German	English
Topp	!Hauptkategorie	Contents
Geografi	**Sachsystematik**	Articles
Historia	Geschichte	**Main topic classifications**
Kultur	Kultur	Culture
Personer	Personen	People
...	...	...

Moreover, Wikipedia maintains administrative and special pages in certain categories under the system namespace that do not constitute main content. These factors add difficulties to the processing of the category graph.

3.2.1 Choosing Semantic Root

Each Wikipedia language edition creates its own category structure with no standard node designated as the root node, nor any standard on where under the root node content-related categories are placed. For instance, as shown in Table 2, content categories are created at the level directly under the root node in the Swedish Wikipedia, two levels below the root in the German Wikipedia, and three levels below the root in the English Wikipedia. Therefore we identify a *semantic root* which constitutes the parent node of the top-most content category nodes. On the other hand, categories are named in their language. Automatically determining the semantic root becomes difficult due to these reasons. Thus the semantic root node needs to be manually identified (shown in boldface in Table 2).

3.2.2 Removing Non-Content Categories

The Wikipedia category structure contains non-content categories which are not useful for our analysis, and indeed would adversely affect the calculation of similarity and the visualization in the later steps. This mainly includes three types of categories: (1) Wikipedia administrative categories, (2) stub categories and (3) list categories. These types of category nodes need to be removed, each of which requires a different approach.

Normally Wikipedia administrative categories are located under the "Wikipedia" namespace, and we can simply drop the categories in this namespace. Stub articles are short articles that need expansion, which are grouped into numerous stub categories. Names of these stub categories usually contain the term "stub", or a translated word with a similar meaning in other languages of Wikipedia. In this way we can easily find and remove them by looking for a particular substring in category names.

List categories, for example "1879 births", "1976 deaths", "History of China by period" and others are convenient for readers to look up articles, but are not useful to include in the final visualization as they are large in number and do not actually contain article content themselves. These categories usually repeat certain keywords in their names, such as "births", "deaths" and "History of". We can record the occurrence of such words in category names. Words that appear frequently in sibling nodes (i.e. under the same parent node in the category graph) are assumed to be part of such list categories, so these nodes are removed.

One characteristic of these list categories is that they share similar category names in sibling categories under the same parent category, for instance, list category "Mammals of Norway", "Mammals of Latvia" and "Mammals of Germany" are placed under the parent "Mammals by country". Given this knowledge we can develop a method to identify list categories by computing similarities of category names. Table 3 shows an example of similarities of sibling categories under the parent category "Aircraft 1950-1959".

Table 3: Category name similarities under category "Aircraft 1950-1959"

Pair of Category Names	Similarity
Civil aircraft 1950-1959 Italian aircraft 1950-1959	0.932
Italian aircraft 1950-1959 Dutch aircraft 1950-1959	0.876
Dutch aircraft 1950-1959 Soviet aircraft 1950-1959	0.899
Soviet aircraft 1950-1959 Military aircraft 1950-1959	0.912
Average Value	0.905

The method for removing list categories works as follows. For each pair of category names we record the occurrence of every character appearing in their names, as well as the number of common characters shared by the pair. A cosine similarity is calculated with these numbers. If the average of the similarities of all sibling categories is greater than a pre-defined threshold (value ranges from 0 to 1, in our case we choose 0.8), then these categories are considered as list categories. This method is language-independent, being applicable also to Asian languages (e.g. Chinese, Japanese and Korean) where the basic linguistic component is a character rather than a word.

3.2.3 Creating a Category Tree

In order to facilitate the analysis process and to simplify our algorithms, we transform the category graph into a simple directed tree. In order to create such a structure, first we apply a breadth-first search that starts from the chosen semantic root, which is derived using the above-mentioned method. The algorithm traverses every node encountered, keeping a list of visited nodes. Loops in the tree are removed by simply eliminating the edge that causes the loop (see Figure 4a), and all parent relationships in multiple-parent nodes are removed except for one (see Figure 4b). Currently the elimination of multiple parents is guided by heuristics that we have devised, but we plan to change this to use cosine similarity based on co-assignment of categories in articles involving the category in question and all its parent categories, which we expect will result in the selection of a more suitable parent category.

3.3 Similarity Calculation

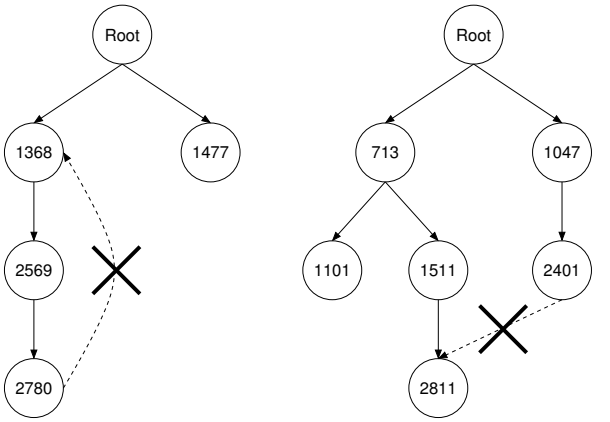
Cosine similarity as used by us is a measure indicating the mutual similarity between a pair of categories. Ususally editors assign an article to multiple categories when the topics of these categories are related to the article's content. We can therefore assume that a pair of categories is more similar to each other if they share many common articles assigned to them. The number of common articles is referred to as *co-occurrence* of category assignments between a pair of categories. A greater number of co-occurrences implies a stronger similarity and vice versa.

$$cos_{i,j} = cos_{j,i} = \frac{\sum_{k=1}^{n} A_k C_{ij}}{\sqrt{\sum_{k=1}^{n} A_k C_i \sum_{k=1}^{n} A_k C_j}} \quad (1)$$

Equation 1 shows the calculation of the cosine similarity [5]. $cos_{i,j}$ represents the cosine similarity of categories C_i and C_j. $A_k C_i$ is the assignment of article A_k to category C_i, and similarly for C_j. $A_k C_{ij}$ is the co-occurrence of article A_k in categories C_i and C_j.

Table 4: Cosine similarity for categories in the English Wikipedia with data from different levels of sub-categories

Pair of Categories	$cos_{i,j}$	1 Level	2 Levels	3 Levels	4 Levels
History – Geography	0.017293	0.001006	0.000081	0.000020	0.000008
History – Culture	0.021874	0.000692	0.000073	0.000022	0.000010
History – Agriculture	0.000000	0.000595	0.000038	0.000009	0.000005
History – Politics	0.024456	0.000715	0.000046	0.000018	0.000010
History – Nature	0.000000	0.001001	0.000045	0.000012	0.000006
History – Technology	0.000000	0.000538	0.000043	0.000009	0.000004
History – Education	0.000000	0.000618	0.000039	0.000012	0.000006
History – Applied sciences	0.000000	0.000705	0.000032	0.000010	0.000008



(a) Eliminating a loop (b) Eliminating multiple parents

Figure 4: Eliminating edges in the category graph (edge indicates "parent category" relationship)

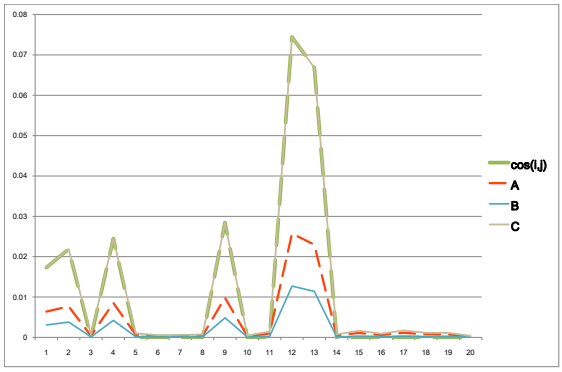


Figure 5: Comparison of different similarity aggregation expressions in English Wikipedia

Table 4 illustrates the cosine similarity values for top level categories of the English Wikipedia. Because of space limitations only a few rows are shown. The table displays the similarities between the category pair in column $cos_{i,j}$, and average values of similarities computed with the inclusion of the sub-categories of these top categories at different depth levels. Since in a larger Wikipedia the category structure is usually more well developed, articles tend to be assigned to sub-categories deep down the category hierarchy instead of directly to top level categories. Thus when comparing top level categories with each other (referred to as *direct similarities*), often no similarity is found (i.e. $cos_{i,j} = 0$). However when subcategories are compared, similarity values are often significantly greater.

3.4 Similarity Aggregation

In order to find similarity values for top level categories it is not sufficient to simply use their direct similarity values, but we should also consider the similarities of their sub-categories. This is because for any pair of categories, their similarity is not only determined by the co-occurrence of the pair. The articles of their sub-categories should also contribute to a certain degree to the relationship of their parents. We therefore define the *aggregated cosine similarity* of a pair of top level categories, which combines both the direct similarity and the similarity from subcategories, as a weighted sum. The weights used were determined empirically by experimentation, using following expressions that combine direct similarity ($cos_{i,j}$) and the average similarities from the corresponding n levels of subcategories ($cos'_{i,j,n}$):

- Expression A: $\frac{1}{3}cos_{i,j} + \frac{2}{3}(cos'_{i,j,1}, \ldots, cos'_{i,j,n})$
- Expression B: $avg(cos_{i,j}, cos'_{i,j,1}, \ldots, cos'_{i,j,n})$
- Expression C: $max(cos_{i,j}, cos'_{i,j,1}, \ldots, cos'_{i,j,n})$

Figure 5 shows the results of experiments on the English Wikipedia. Firstly, the values of the direct similarity $cos(i,j)$ and Expression C are not desirable because they magnify the differences between categories with zero and high values. We also applied tests on Wikipedia in other languages. Standard deviations of values obtained with different tests are shown in Table 5. Expression B has the lowest standard deviation, which means it can minimize the effect of extreme values across categories, while it retains the characteristics of the values as shown in Figure 5. As a result, we suggest to compute the similarities of top level categories as follows:

$$ac_{i,j,n} = avg(cos_{i,j}, cos'_{i,j,1}, ..., cos'_{i,j,n}) \quad (2)$$

The analysis and visualization of Wikipedia described in the forthcoming sections is based on this equation.

4. VISUALIZATION METHOD

This section discusses a new method for visualizing wikis in a form resembling a geographic map. The map displays a "virtual territory", with no correspondence to any real geographic area. We merely use the representational form of a geographic map to present an easily understandable visualization. Categories are represented as areas in the map, with sub-divisions into sub-areas corresponding to sub-categories. Our method uses the similarity of

Table 5: Standard deviations of different similarity aggregation expressions in different Wikipedia language editions

Expression	Danish	Chinese	Swedish	German	English
$cos_{i,j}$	0.015995	0.004116	0.013223	0.011817	0.010498
Expression A	0.006372	0.001412	0.004425	0.006040	0.003513
Expression B	0.003932	0.000840	0.002659	0.002083	0.001761
Expression C	0.016071	0.004066	0.013147	0.013397	0.010286

wiki categories to place categories with closer relationship nearer to each other. The principles and method described here are generic enough to apply to all Wikipedia language editions, as well as to other wikis based on the Wikipedia wiki engine (MediaWiki), and should also be applicable to other wiki systems that have a categorising feature. As a case study, we present our visualization of the English and German Wikipedias.

4.1 Preliminary Layout

Creating a rough layout of wiki categories is the first step in generating a map-like visualization. This *preliminary layout* contains approximate positions and estimated sizes of categories in the final visualization. We use a bottom-up approach to create this layout, proceeding upwards level by level. From our experience with Wikipedia data we found that up to three levels of categories could be well represented without resulting in too many or too small subdivisions. For example top-level category "Science" contains subcategory "Mathematics", which contains sub-sub-category "Geometry". To follow the metaphor of a geographic map, these three levels can be thought of as three levels of political regions namely countries, provinces (or states), and counties.

The bottom-up algorithm starts at the lowest defined category level, and iterates over every level until it reaches the top level. Sub-categories in the current level are placed using a force-directed spring layout algorithm [7]. Similarities between pairs of categories are fed into the spring algorithm, acting as "forces" between categories. The layout algorithm adjusts the positions of categories until they are stable and forces are balanced. During the execution of the algorithm, a given category level settles positions of its categories and results in a bounding box that contains all its category nodes. After the algorithm has finished processing a given category level, the next higher category level combines the layouts from the lower level and adjusts their positions using their mutual similarity values. The algorithm continues to iterate until it reaches the top level and layouts of the top level categories have been determined.

The layout at this stage may contain overlaps because the force-directed algorithm considers only positions of node points and not areas when balancing overall forces. Thus an overlap removal algorithm is applied as a final step. In our method we use the Force Transfer Algorithm (FTA) [6]. As we mentioned in Section 2, FTA can remove overlaps while keeping the layout compact.

4.2 Finalizing the Map-like Visualization

Similar to Skupin's map-like visualization of a document corpus [15], we also make use of hexagons to create a "map" for wikis. Factors such as irregular shapes, various line length, colours used, etc. are important in achieving a visualization that looks realistic and resembles a geographic map. Using hexagons has two advantages: (1) hexagons tile a surface completely, and (2) border lines of areas tiled by hexagons have a natural-looking irregular appearance. Thus hexagons are a suitable choice as the basic tiling shape.

Our algorithm creates a hexagon lattice in memory. Its main purpose is to allocate hexagons in the lattice to categories according to the preliminary layout, resulting in regions that represent their categories. The number of hexagons occupied by a category is proportional to the number of sub-categories and articles it contains. As the numbers of articles and sub-categories can have extreme variations in some Wikipedia language editions, a logarithmic scale is applied to area size in order to reduce the occurrence of very large category regions, and also to make small categories more noticeable. The assignment of hexagons is performed randomly from a given starting hexagon (the area's *pivot point*). A data structure maintains a list of "territory" (i.e. occupied hexagons) of the current category and randomly selects unused neighbouring hexagons for assignment, as illustrated in Figure 6 (numbers correspond to the order of hexagon selection). Through experimentation we discovered that this form of random assignment produces regions that look most similar to those in a geographic map. In order to create reproducible output, however, we control randomness throughout our program by using a fixed random seed to the pseudo-random number generator. Thus the same input data will always result in the same visualization, and make different visualizations comparable with each other.

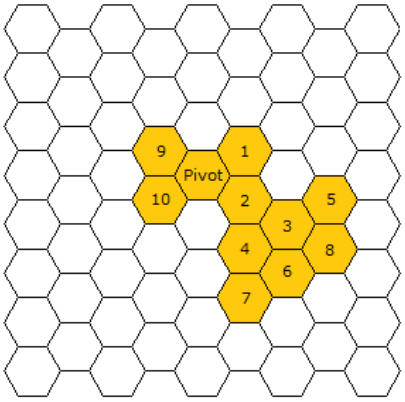


Figure 6: Example of random hexagon assignment for a category

Colour is one of the primary visual elements that a reader perceives in a visualization. Topographic maps often employ colour to represent elevation. We use it to represent the total number of articles of a category. This allows users to quickly spot large and small categories, and to perceive their distribution throughout the visualization. The colour scheme we employ is similar to that of topographic maps: a darker colour represents a higher value (in our case meaning a higher number of articles).

Finally text labels are added. Text labelling can be problematic when the density of labels to be placed in a given area is high, such as is the case in our visualization of Wikipedia. Geographic maps often are created using manual placement of text label to avoid overlaps and produce an optimal appearance. However, given the size of Wikipedia datasets this approach is clearly infeasible. Therefore we place text labels automatically. Where necessary, we reduce font sizes of label text to enable labels to remain inside the area of their respective categories, with a minimum allowed font

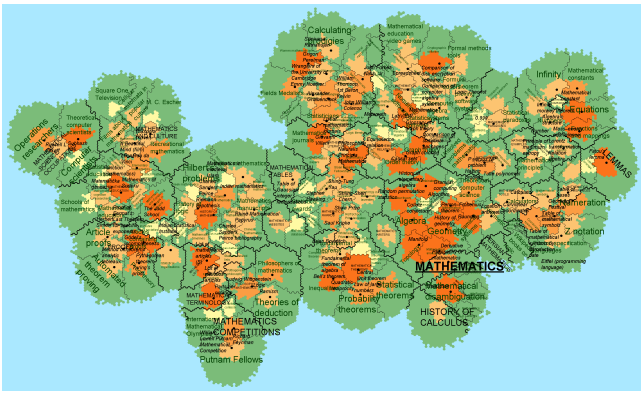


Figure 7: Highlight of category "Mathematics" in the English Wikipedia, showing sub-categories and their relative sizes

size to ensure readability (meaning that in a few cases of long labels in very small areas labels could protrude from their areas). Text styles are used to distinguish label hierarchies. For example top level categories are shown as in "**SCIENCE**", second level categories as in "**MATHEMATICS**", and third level categories as in "Geometry".

5. DISCUSSION

As mentioned in Section 3, we have applied our visualization method to five Wikipedia language editions: Danish, Chinese, Swedish, German and English, in order to informally assess the visualization result. The subsections below mainly focus on the two largest Wikipedias, English and German, and discuss the use of our visualizations in different areas.

5.1 Understanding a Wiki's Category Composition

Wiki content is organized by categories, and to understand the current topic distribution in a wiki we can look at the distribution of articles over the existing categories. Doing so manually would require inspecting many categories and sub-categories one by one. Our map-like visualization provides a much more easily perceivable category composition of a given wiki. For example Figure 7 shows a highlight of category "Mathematics" in the English Wikipedia. This figure shows at a glance both the sub-categories, such as "LEMMAS" and "HISTORY OF CALCULUS", as well as sub-sub-categories, such as "Algebra" and "Geometry". Relative sizes can be easily perceived, which reflects different popularity of different topic areas among authors, which is further amplified by different area colours. For instance, category "Geometry" is shown with two colour levels darker than nearby category "Arithmetic", reflecting its much larger number of articles. Overall about half of all categories are in the darkest area colour, indicating a high level of content maturity, but sub-sub-categories coloured in the lightest colour can be seen in almost all sub-categories, indicating a need for further content development. This type of overview allows editors to quickly perceive the current content collection of a wiki, such as when identifying large topic areas for potential division into sub-categories, or for identifying relatively under-represented topics that deserve more attention.

5.2 Representing Category Similarity

Category similarity measures the extent to which different categories in a wiki are related by content. Similarity values can provide answers to questions such as: "Which categories are more related to Belief?", or "Are categories Society and History closely related?". While similarity values in themselves can provide an answer, yet in their numeric form it is difficult to perceive multiple mutual relations, and this becomes more difficult as the number of categories involved increases.

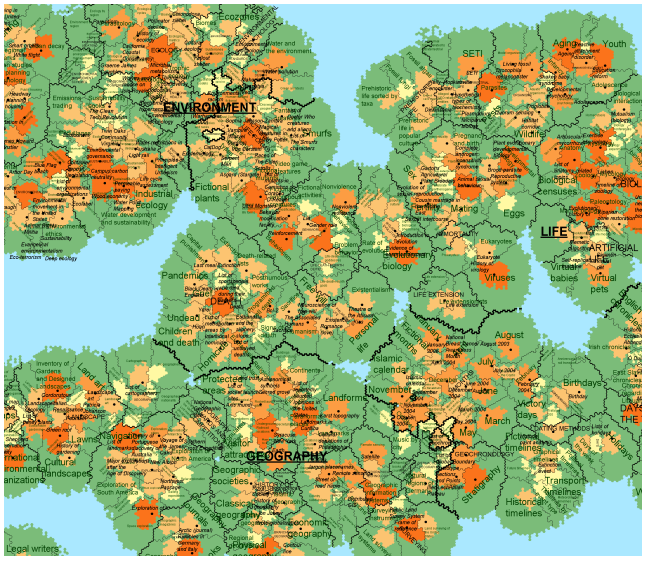


Figure 8: Cluster of related categories "Environment", "Life" and "Geography" placed in close proximity of one another

However, when viewing the visualization, proximities clearly express the relationships between categories. An example of this is shown in Figure 8. It shows the three neighbouring top-level categories "Environment", "Life" and "Geography". The content of these categories is closely related to each other, which is expressed in the visualization by placing them adjacent to each other. On the other hand, other much less related top-level categories such as "Arts" and "Belief" are placed much further away (not pictured in Figure 8). Thus the visualization summarises the overall relationship of these categories into visual proximity. This form of visualization provides a straightforward way to discover relations among topic areas in a wiki.

5.3 Overview of Multiple Wikis

A map-like visualization effectively serves as a "world map" of a wiki. It provides a first impression of the overall situation of that wiki, similar to how a real world map gives an overview of the distribution of land and sea, the relative sizes and positions of continents, etc. Using visualizations of multiple Wikipedia language editions makes comparisons possible, just like using several political maps helps compare aspects of different physical countries. For instance, Figures 9 and 10 give an overview of the German and English Wikipedia editions, respectively.

The first impression we get from these visualizations is a difference in colour: the German Wikipedia is more strongly dominated by dark orange area colours than the English Wikipedia. This would indicate that the categories contain more content. Given the background knowledge of the size of these Wikipedias (the English Wikipedia having about three times as many articles as the German one) this would indicate that the category hierarchy in the German Wikipedia is not as finely divided as is the English one. Table 6 presents the relevant figures: although the number of top-level categories in the German and English Wikipedias does not

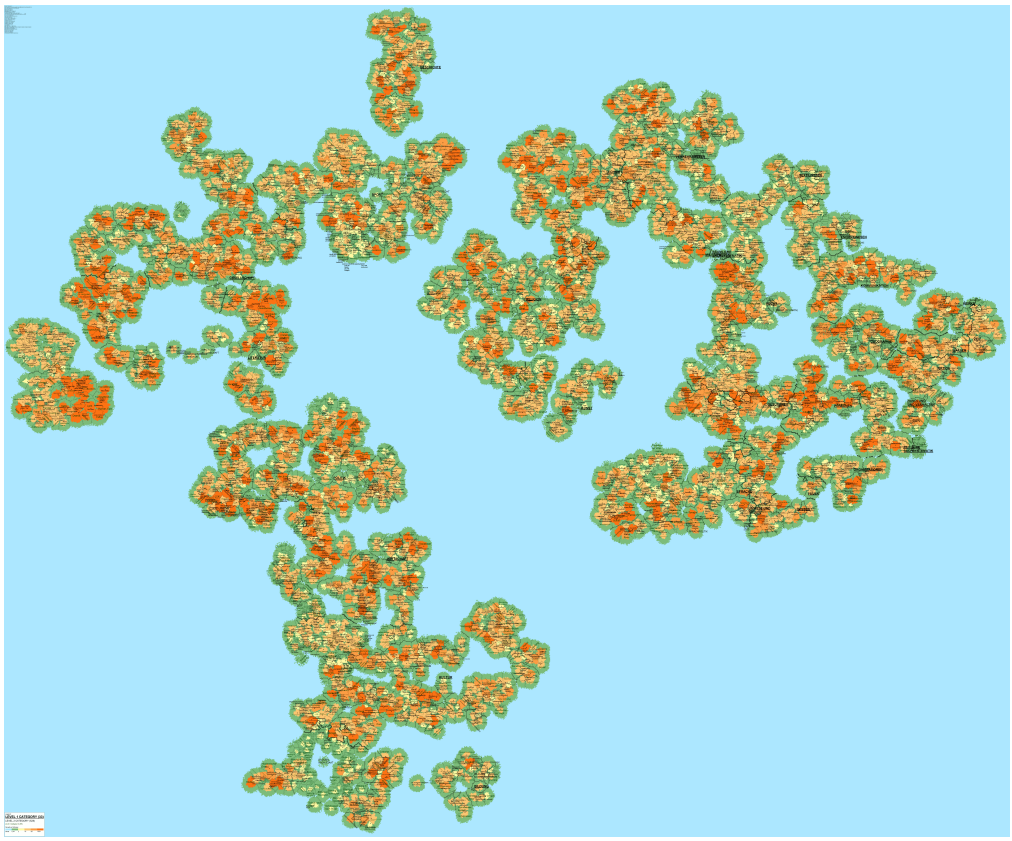


Figure 9: Overview map of the German Wikipedia

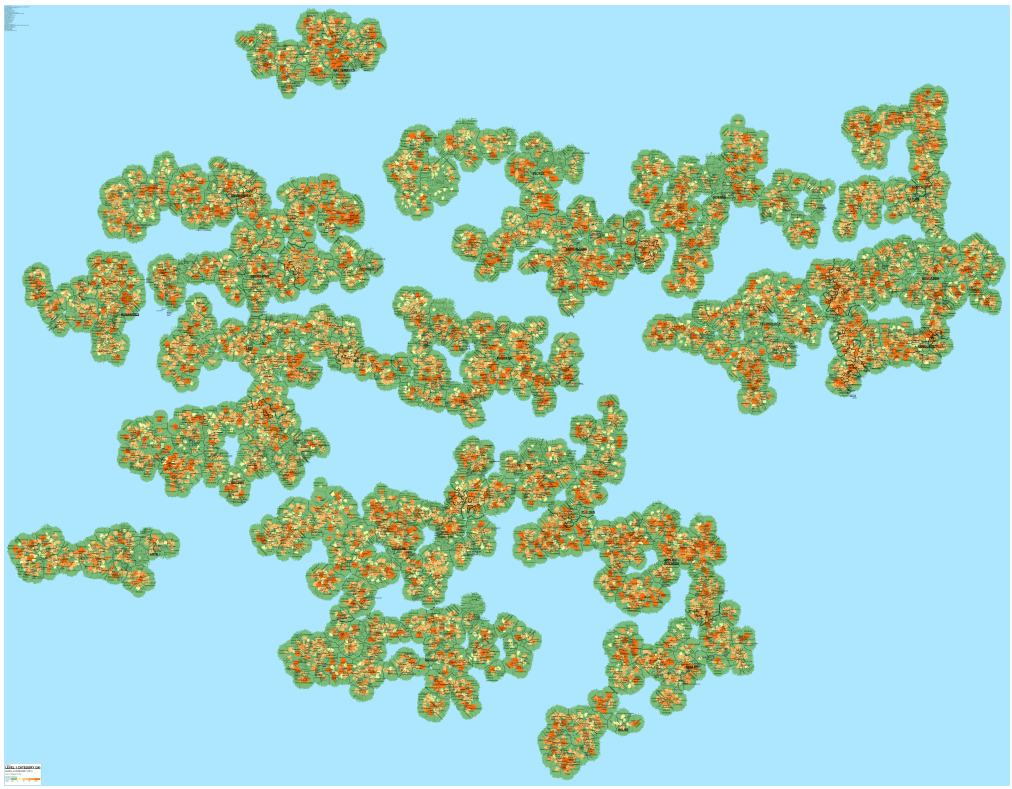


Figure 10: Overview map of the English Wikipedia

Table 6: Statistics of German and English Wikipedia

	German	English
No. of top-level categories	28	24
No. of all categories	84,161	602,141
No. of all articles	1,055,243	3,411,491
Avg. articles per category	12.5	5.7

Table 7: Statistics of category "Science" in the Danish, Swedish and Chinese Wikipedia with 2 levels depth

	Danish	Swedish	Chinese
No. of sub-categories	21	30	34
No. of sub-sub-categories	65	186	105
Total no. of sub-categories	86	216	139
Total no. of articles	2382	7867	1563
Average articles per category	27.70	36.42	11.25

differ strongly from one another, the total number of categories varies greatly. In the English Wikipedia there are about 7 times as many categories as in the German one, but only about 3 times as many articles. Correspondingly the average number of articles per category in the German Wikipedia is more than twice that of the English Wikipedia. Clearly the English Wikipedia's user community favours dividing their content into finer topic sub-divisions, with the effect that each category has relatively fewer articles than in other language Wikipedias, hence the relative lack of the darker orange colours standing for categories with more article content.

Another difference that can be perceived between these two language Wikipedias is that in the German Wikipedia there is a greater rift between clusters of categories on the left and right, with more "sea" area occupying the divide. In the English Wikipedia, however, there are not such great "sea" areas dividing categories. This may reflect the fact that in the English Wikipedia many more articles are assigned across multiple categories belonging to different knowledge domains, i.e. different top-level categories, which in turn would indicate that articles are more integrated with each other and with other knowledge areas, whereas in contrast articles in the German Wikipedia are somewhat less integrated and more independent of each other. Further investigation would be warranted to confirm whether this hypothesis is true, but we believe the value of a map-like visualization like this is in allowing exploration of the wiki content which results in observations such as the above.

5.4 Comparison of Specific Topic Areas across Multiple Wikis

Besides comparing entire wikis, a map-like visualization can be used to compare the same or similar topic areas across wikis. For example, we can compare characteristics of a topic area in different Wikipedia language editions to discover the relative maturity of the same topic across these wikis. Table 7 shows some basic statistics related to categories and articles under the "Science" top-level category in the Danish, Swedish and Chinese Wikipedias. These figures show that this category is the most developed in the Swedish Wikipedia, in terms of all of: total number of sub-categories, total number of articles, and average number of articles per category.

The corresponding map-like visualizations of this category are displayed in Figure 11 (applying the same scaling ratio for resizing all three images to allow comparability). These visualizations first demonstrate the size of content among the different wikis. The overall area occupied by this category is the largest in the Swedish Wikipedia, followed by the Chinese and Danish ones. This matches the figures shown in Table 7. Besides, the average number of articles in the Chinese Wikipedia is actually the lowest, although it has more sub-categories under the "Science" category than either of the other two wikis. The colours in our visualization, related to the *density* of articles in categories, accurately reflect this fact: many areas in the Danish and the Swedish visualizations are displayed in darker colours, whereas most Chinese sub-category regions are displayed in lighter colours.

6. CONCLUSIONS

We have presented a new method for visualizing a wiki in a form resembling a geographic map, and shown applications of these visualizations by examples taken from Danish, Swedish, Chinese, German and English Wikipedia language editions. Visualizations such as these have the potential to reveal in a readily perceivable way much information contained within large wiki article collections.

A number of different kinds of visualizations of Wikipedia data exist. Some of these are also able to provide overviews of a wiki. However, one of the unique features of our work is that we represent the abstract topic coverage of Wikipedia in the form of a (virtual) geographic map, which is easily understandable by a wide range of users. Secondly, our visualization displays an overview of the entire Wikipedia, whereas some other visualization tools focus only on the evolution of articles or the behaviour of authors.

Our research is still a work in progress and we are actively moving this work forward in several directions. Performance is of critical importance in processing large volumes of data, which we are working on improving to allow us to more easily visualize the largest Wikipedia language editions. Moreover, we are also planning to add more map elements, such as roads between cities representing significant linkages between the corresponding articles. This will enrich our visualization by increasing the information communicated by it.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge the support for this research from the Macau Special Administrative Region – Science and Technology Development Fund (grant number 021/2011/A).

8. REFERENCES

[1] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001.

[2] B. Hecht and D. Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the Fourth International Conference on Communities and Technologies*, C&T '09.

[3] B. Hecht and D. Gergle. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 291–300, New York, NY, USA, 2010. ACM.

[4] B. J. Hecht and D. Gergle. On the "localness" of user-generated content. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 229–232, New York, NY, USA, 2010. ACM.

[5] T. Holloway, M. Bozicevic, and K. Börner. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, 12-3:30–40, 2006.

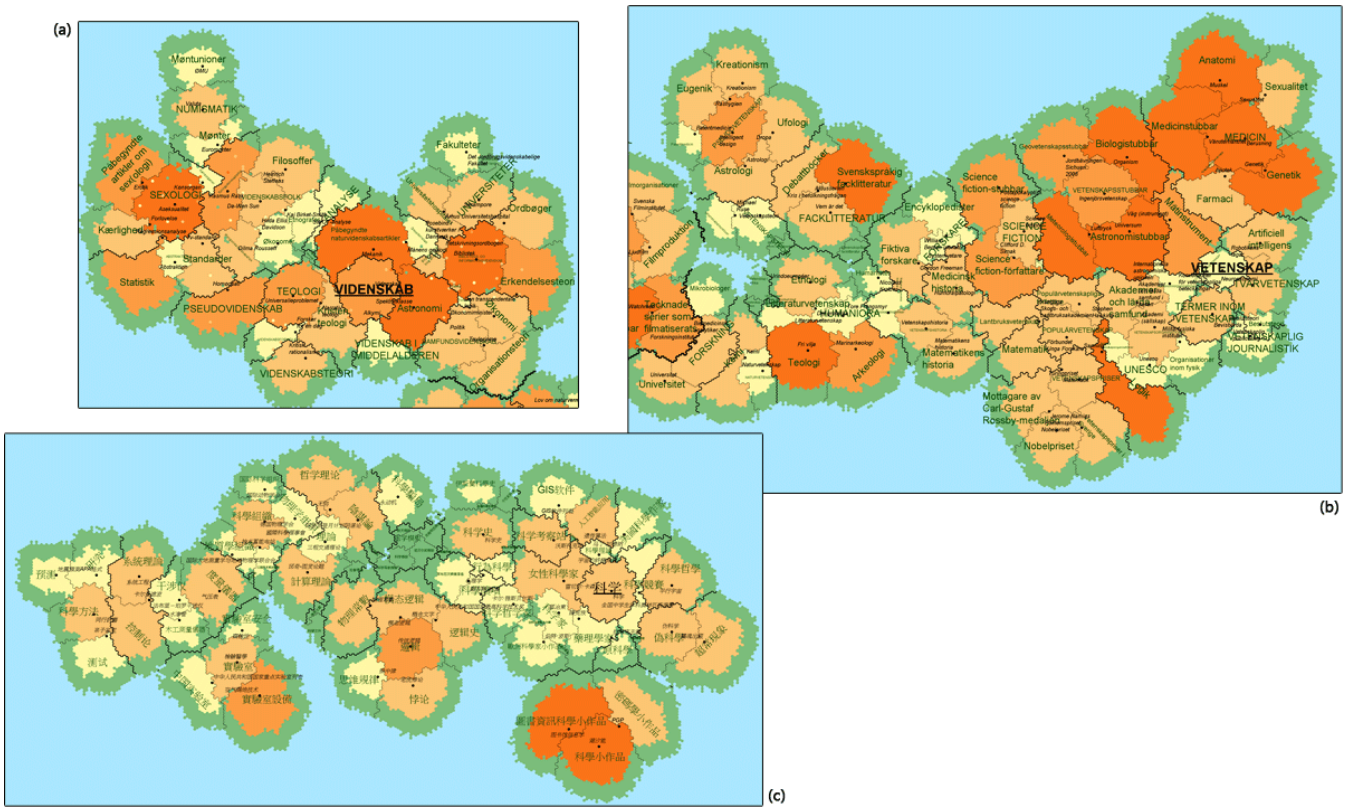


Figure 11: Comparison of category “Science” in Wikipedia: (a) Danish (b) Swedish (c) Chinese

[6] X. Huang and W. Lai. Force-transfer: A new approach to removing overlapping nodes in graph layout. In *The Twenty-Fifth Australasian Computer Science Conference (ACSC2003)*, pages 349–358, 2003.

[7] T. Kamada and K. S. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, Apr. 1989.

[8] A. Kittur, E. H. Chi, and B. Suh. What’s in Wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, CHI ’09, pages 1509–1512, New York, NY, USA, 2009. ACM.

[9] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences, 2001.

[10] Y. H. Li and A. K. Jain. Classification of text documents. *The Computer Journal*, 41(8):537–546, 1998.

[11] M. D. Lieberman. You are where you edit: locating Wikipedia users through edit histories. In *Proceedings of Third International AAAI Conference on Weblogs and Social Media*, pages 106–113, 2009.

[12] K. Misue, P. Eades, W. Lai, and K. Sugiyama. Layout adjustment and the mental map. *Journal of Visual Languages & Computing*, 6-2:183–210, 1995.

[13] D. O’Leary. Wikis: ‘From each according to his knowledge’. *Computer*, 41(2):34 –41, Feb. 2008.

[14] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[15] A. Skupin. The world of geography: Visualizing a knowledge domain with cartographic means. In *Proceedings of the National Academy of Sciences*, volume 101 (Suppl. 1), pages 5274–5278, 2004.

[16] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 575–582, 2004.

[17] M. Wattenberg, F. B. Viégas, and K. Hollenbach. Visualizing activity on Wikipedia with Chromograms. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part II*, INTERACT’07, pages 272–287, Berlin, Heidelberg, 2007. Springer-Verlag.

[18] J. Yu, J. A. Thom, and A. Tam. Ontology evaluation using Wikipedia categories for browsing. In *Proceedings of the Sixteenth ACM conference on Conference on Information and Knowledge Management (CIKM ’07)*, pages 223–232, 2007.

[19] T. Zesch and I. Gurevych. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, pages 1–8, 2007.