Future Generation Computer Systems 31 (2014) 120–133

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Visualizing large-scale human collaboration in Wikipedia

Robert P. Biuk-Aghai *, Cheong-Iao Pang, Yain-Whar Si

Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Av. Padre Tomas Pereira, Taipa, Macau S.A.R., China

HIGHLIGHTS

- A novel method for analysis and visualization of large wikis such as Wikipedia.
- Visualization of a wiki in a form similar to a geographic map.
- Analyzed and visualized English, German, Chinese, Swedish and Danish Wikipedia.
- Significant co-author count differences between different language Wikipedias.
- Superior over text data in usability, accuracy, speed and user preference.

ARTICLE INFO

Article history:
Received 9 January 2012
Received in revised form
10 November 2012
Accepted 6 April 2013
Available online 23 April 2013

Keywords:
Visualization of collaborative processes & applications
Wikipedia
Information visualization
Category
Co-authoring

ABSTRACT

Volunteer-driven large-scale human-to-human collaboration has become common in the Web 2.0 era. Wikipedia is one of the foremost examples of such large-scale collaboration, involving millions of authors writing millions of articles on a wide range of subjects. The collaboration on some popular articles numbers hundreds or even thousands of co-authors. We have analyzed the co-authoring across entire Wikipedias in different languages and have found it to follow a geometric distribution in all the language editions we studied. In order to better understand the distribution of co-author counts across different topics, we have aggregated content by category and visualized it in a form resembling a geographic map. The visualizations produced show that there are significant differences of co-author counts across different topics in all the Wikipedia language editions we visualized. In this article we describe our analysis and visualization method and present the results of applying our method to the English, German, Chinese, Swedish and Danish Wikipedias. We have evaluated our visualization against textual data and found it to be superior in usability, accuracy, speed and user preference.

1. Introduction

The emergence of Web 2.0 technologies in recent years has made human-to-human collaboration on unprecedented scales not only possible but a reality. One of the best-known examples of world-wide large-scale collaboration is Wikipedia, "the free encyclopedia that anyone can edit" (Wikipedia's own slogan) [1]. Wikipedia has great value that has not yet been fully researched. Past research on Wikipedia has focused on both a *micro-level* (e.g. [2,3]) and a *macro-level* of analysis (e.g. [4–7]). 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 and aims 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 language editions, and by implication to discover 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, for example intra-organizational wikis.

The technology underlying Wikipedia is relatively simple: a wiki engine (MediaWiki) implemented in PHP on a web server which most users access through a web browser, and primarily making use of three main functions: searching for, reading and editing articles. Other functions, used to a much lesser extent by common users, are asynchronous discussion of articles, viewing the revision history of an article, comparing revisions to find out what has changed between them, undoing specific revisions, and a few others. Wikipedia administrators have additional privileges, allowing them to protect articles (making them read-only), moving (renaming) articles, deleting articles entirely, blocking users, and other administrative/maintenance functions.

The Wikipedia user base is large and broad: the English Wikipedia edition alone counted about 17.8 million registered

* Corresponding author. Tel.: +853 83974375.
E-mail addresses: robertb@umac.mo (R.P. Biuk-Aghai), inbox@patrickpang.net (C.-I. Pang), fstasp@umac.mo (Y.-W. Si).

0167-739X/$ – see front matter
http://dx.doi.org/10.1016/j.future.2013.04.001

Table 1
Wikipedia user statistics, as at 8 Nov 2012 (active user % is relative to all registered users, admin user % is relative to active users).

Language	Users				
	Registered	Active		Admins	
		Total	(%)	Total	(%)
English	17,813,716	132,800	0.7	1462	1.1
German	1,535,302	21,649	1.4	267	1.2
Chinese	1,316,773	6,994	0.5	78	1.1
Swedish	299,093	3,136	1.0	88	2.8
Danish	171,699	1,155	0.7	37	3.2

users in November 2012, out of which 132,800 (0.7%) are considered "active" users (meaning that they have performed some action within the past 30 days). A small portion of these registered users are site administrators, under 1500 (about 1% of active users) in the case of the English Wikipedia. An overview of user statistics for a few selected Wikipedia language editions that we have studied is shown in Table 1. We selected these Wikipedia language editions mainly for the practical reason that we understand these languages (which is required for interpreting the visualized result), but also to give us a selection of very large (English), medium-sized (German, Chinese) and small (Swedish, Danish) Wikipedias.

Wikipedia content is user-contributed, meaning that end-users can add to, modify and delete content in Wikipedia articles. They can also write entirely new articles and link these to other articles. To better organize content Wikipedia has a hierarchical category system, and any given article can be marked as belonging to any number of categories. For instance in the English Wikipedia (as of January 2012), article "Wiki" is assigned to category "Wikis" (plus five other categories), which in turn has parent category "World Wide Web" (plus four other parent categories), which in turn has parent category "Digital Media" (plus six other parent categories), and so on. The same as with articles, categories are also user-contributed: users can create new categories, assign categories to parent categories, assign articles to categories, and change existing article-to-category and category-to-category assignments. The result is an organically evolving category system that reflects the current needs of the user-contributor community. One of the implications of such an open editing process is that it may result in different granularity of the category hierarchy. Table 2 shows the numbers of articles and categories of the five Wikipedia language editions we have analyzed (these counts include all articles and categories, including non-content ones that we later remove). The absolute numbers of articles and categories differs significantly in these different language editions, but so does the average number of articles per category (the right-most column in Table 2) which indicates the granularity of the category hierarchy. In four of the five analyzed Wikipedia language editions the number of articles per category ranges between about 4 and 8, but in the German Wikipedia there are on average 17.7 articles per category, suggesting a much coarser category hierarchy granularity. As documented on Wikipedia itself, the German edition of Wikipedia differs from other editions: "Compared to the English Wikipedia, the German edition tends to be more selective in its coverage" and "Categories are usually introduced only for a minimum of ten entries and are not always subdivided even for larger numbers of items,"[1] which explains this difference in the articles per category statistics. In fact, the absolute number of categories in the German Wikipedia is even smaller than that in each of the Chinese and Swedish Wikipedias although the number of articles is significantly larger. Different language communities clearly have different standards as to how fine-grained they believe their category hierarchies should be.

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

Table 2
Sizes of Wikipedia language editions studied (database dump of January 2011).

Language	No. of articles	No. of categories	Art./Cat.
English	3,411,491	602,141	5.7
German	1,217,553	68,677	17.7
Chinese	352,562	82,639	4.3
Swedish	393,504	82,039	4.8
Danish	147,576	19,193	7.7

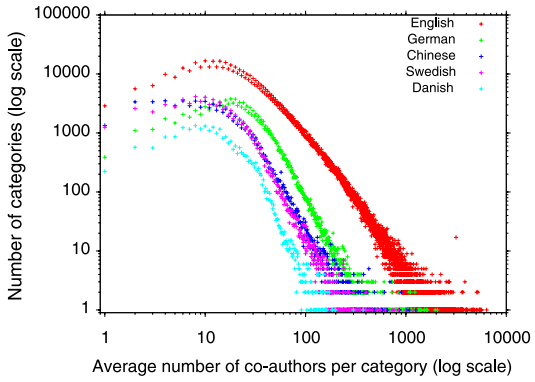


Fig. 1. Distribution of the average number of co-authors per category in English, German, Chinese, Swedish and Danish Wikipedia.

Wikipedia is not only user-contributed, but as a direct result of its openness the number of contributors that get involved in editing a given article can also be very large. We have analyzed this number of co-authors and for each category calculated the average number of distinct co-authors of all articles assigned to that category. This average count of co-authors per category varies dramatically between categories. For example, in the English Wikipedia there are 15 categories, each of which has an average number of co-authors of over 5000. On the other hand there are over 100,000 categories, each of which has an average number of co-authors of 10 or fewer. The distribution of average number of co-authors per category in the five Wikipedia language editions we analyzed is plotted in Fig. 1. Interestingly, despite all the differences in scale and category hierarchy granularity among the different language Wikipedias, their curves have essentially the same shape. We determined goodness of fit using the Anderson–Darling test and found the data from all five language editions to follow a geometric distribution, with p ranging between 0.03 and 0.05 in the different languages.

However, this distribution of the average number of co-authors per category does not reveal *where* the differences lie—which categories attract the most co-authors to their articles, and which the fewest. This may also differ between different Wikipedia language editions, as the top-10 list of categories with highest co-author count shown in Table 3 indicates politics and religion feature strongly in the English Wikipedia, whereas in the German Wikipedia it is art and society that feature strongly, with some sports and television appearing in both top-10 lists. We also do not know if similar co-author counts cluster together by topic, i.e. whether categories that belong to the same parent category also have similarly high co-author counts. This information is difficult to obtain as topic clusters are hard to determine given the large number of parent categories that a given category may belong to.

We have devised a method for analyzing the category hierarchy to determine which major parent category a given category should belong to. This allows us to aggregate co-author counts from individual categories recursively up to their ancestor until the top of the category hierarchy. Doing so reveals which categories at the highest level are the most collaborative, and which the least. We have then used the output of this analysis to visualize the

Table 3
Top 10 most co-authored categories in English and German Wikipedias within the top 3 levels of categories (English translations of German category names in brackets).

Rank	English	German
1	Superpowers	Kunst (Arts)
2	Christians	Brücke (Bridge)
3	Sports	Die Simpsons (The Simpsons)
4	Hindus	Architekt (Architect)
5	Foods	Archäologie (Archaeologie)
6	Television	Gesellschaftliche Schicht (Social class)
7	World	Kunst nach Staat (Arts by country)
8	Religious sceptics	Konfuzius (Confucius)
9	Places	Fußballtorhüter (Football goalkeeper)
10	Existentialism	Harry Potter (Harry Potter)

average numbers of co-authors across the first three levels of the category hierarchy for the entire Wikipedia of a given language. This visualization reveals interesting differences in the distribution of co-authoring over the Wikipedias we studied.

The remainder of this article is organized as follows: the next section briefly presents related work. Section 3 then introduces our analysis and visualization method. Section 4 presents the results of applying our method to actual Wikipedia data, followed by Section 5 which evaluates our visualization. In Section 6 we discuss applications of our visualization method and make conclusions in Section 7.

2. Related work

Wikis in general, and Wikipedia in particular, have experienced dramatic growth over the past decade both in size and value. Consequently they have become the focus of research by numerous researchers in different fields worldwide. This section gives a brief overview of pertinent research.

2.1. Wiki category pre-processing

Wikipedia, the wiki that is the focus of our work, has a category organization that is similar to a tree structure. However, because the creation and maintenance of the category hierarchy is a manual task performed by Wikipedia users some cases such as multiple parents, loops, and other anomalies occur. Indeed the category hierarchy of Wikipedia can be classified as a kind of *directed 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 hierarchy 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 [8]. 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 use a depth-first search to traverse the category graph. Whenever they detect a cycle among the nodes of the same level, they simply remove one of the links on that level to eliminate the cycle [9].

Our pre-processing of Wikipedia categories uses the same approach as that of Yu et al. to remove multiple parents, but unlike Zesch and Gurevych we use a breadth-first search for detecting cycles, which results in a tree with nodes connected to the root through the shortest possible distance.

2.2. Category similarity calculation

Our visualization method creates an overview of a wiki primarily based on the relationship among categories. Once relationships are known for all pairs of categories we can calculate individual positions of all categories in the drawing plane by laying out category nodes using an approach such as force-directed layout [10]. The relationship of categories can be visually represented by their proximity, i.e. more similar categories are placed closer to each other.

Holloway et al. introduced a method for computing similarities among wiki categories using the number of *co-assignments* of the same categories in articles [4]. 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 these connections (i.e. co-assignments) imply a stronger relationship between categories. Cosine similarity has been used for a long time in computing similarity between articles linked by identical keywords [11–13], but it is innovative to apply the method for calculating category relationships. This is also the approach we adopted.

Szymański employed a somewhat different approach, using links between articles to infer category links [14]. Semantic similarity of a group of articles is calculated and is used to add weights to existing category links expressing the strength of their similarity. It can also be used to generate new links between categories that the group of similar articles link to and that were previously unconnected and can even be used to generate entire new categories corresponding to the concepts represented by that group of articles.

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 which visualizes the evolution of an article over successive revisions [2]. 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 [4]. Some types of visualization focus on analyzing users' activities and authorship. Wattenberg et al. created an application called Chromograms [15] which displays operations performed on the content of Wikipedia, such as spell-checking, writing new content, reverting changes, etc. Harrison created a number of visualizations for Wikipedia including its top 50 articles, the structure of interconnections between Wikipedia

category pages, and visualization of the full Wikipedia graph.[2] Other interesting projects on Wikipedia include visualization of frequent words in English Wikipedia, real-time visualization of Wikipedia edits, visualization of article growth in the Cebuano Wikipedia, visualization of the number of articles at a chosen date, visualization of the differences between the category structure of the Universal Decimal Classification (UDC) system and Wikipedia, and visualization of the edit history of a page.[3]

Our visualization is a case of whole-wiki visualization and is unique as it is the only one that visualizes co-author counts.

2.4. Map-like visualization

Most people understand geographic maps easily, thanks to early exposure to maps in school. Even among pre-school children essential mapping abilities are well developed [16]. 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.

Data visualized in map form is usually multi-dimensional. For example, for each of the 24 top-level categories in the English Wikipedia there are 23 relationships to other categories, meaning that each data point in this set is at least 23-dimensional, with other attributes such as the number of articles, the number of co-authors, etc. adding further dimensions. To represent such a dataset in the map form means transforming this high-dimensional data to a small number of dimensions. For example if reduced to four dimensions these can be represented as x and y coordinates of a shape, its area size, and colour.

Skupin presented a method that produces a map-like visualization for a knowledge domain based on the *Self-Organizing Map (SOM)* framework [17]. 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 [18]. Skupin's method was novel to apply the SOM framework 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. However, Skupin's approach produces a single contiguous area (''continent''), rather than a set of areas separated by empty space (''oceans''), which limits the distance that unrelated parts of a knowledge domain can have in the visualization, making some parts look more related than they should be.

To reduce the number of dimensions, other methods employed include principal component analysis (PCA) and multi-dimensional scaling (MDS). Principal component analysis uses an orthogonal linear transformation of a set of variables into a (usually smaller) set of variables [19]. It identifies the variables that account for the largest variance in the data and thus are the most informative in expressing characteristics of that data. The use of PCA in visualization is mainly in the pre-processing of data (e.g. [20,21]), although Müller et al. have applied it to enhance the visualization process itself [22].

Multi-dimensional scaling is an approach for mapping high-dimensional data into two- or three-dimensional space, so that the data can be visualized. Input data consists of distance values for pairs of data items. The desired number of dimensions is decided in advance (usually two or three, for mapping to coordinates in a drawing space), and the data is mapped using one of a number of available MDS algorithms. The outputs can be used as coordinates for producing a visualization. Many examples of MDS in data and information visualization exist, such as [23–25] to name only a few.

Our presented approach is unique in utilizing similarities between data objects to create a planar layout in which proximity represents similarity.

2.5. Multi-language and cross-language Wikipedia studies

In the domain of information retrieval Wikipedia has been welcomed as a resource for conducting experimental studies. An approach for cross-language information retrieval that is an extension of the explicit semantic analysis method was separately proposed by Sorg and Cimiano [26] and Potthast et al. [27]. They have applied this to Wikipedia articles and evaluated the performance of their method for cross-lingual information retrieval. In related work Sorg and Cimiano induced cross-language links between articles in different language editions of Wikipedia [28].

Hautasaari et al. present tools to facilitate the work of the Wikipedia translation communities that translate Wikipedia content across language editions [29].

The Wikipedia category network has been used by Nastase et al. to relate concepts during the creation of a large multi-lingual concept network based on Wikipedia content [30].

Our visualization of entire Wikipedias has a somewhat different focus and can be used to visually compare various aspects of different language editions.

3. Analysis and visualization method

Our analysis and visualization method consists of two parts, respectively, for analysis and visualization of Wikipedia data. The analysis part processes the data in preparation for the visualization part which transforms the data into a graphical form. The pre-processing immediately precedes the visualization; thus if the category data is changed in any way then running the pre-processing and visualization process anew will produce a visualization that reflects this change.

The visualization produced resembles a geographic map. As discussed above, most people know how to read geographic maps and intuitively relate to them. Thus this form of presentation achieves ease of use and obviates the need to learn how to read the visualization. Fig. 2 shows our analysis and visualization method in outline. It consists of the following steps.

Step I.1: *Select a semantic root*. Wikipedia does not have a standard name for the root node of the category hierarchy, nor any standard for where under the root node content-related categories are placed. For example, in the Swedish Wikipedia, content categories such as "History" ("Historia" in Swedish) are placed directly under the root node ("Topp"), whereas in the German Wikipedia the equivalent category ("Geschichte") is placed two levels below the root, and in the English Wikipedia three levels below the root (see Table 4). Thus it is necessary to manually inspect the category hierarchy and to identify the node under which content categories are placed, which we designate as the *semantic root*.

Step I.2: *Remove non-content categories*. The Wikipedia category hierarchy 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.

[2] http://www.chrisharrison.net/index.php/Visualizations/.
[3] http://infodisiac.com/Wikimedia/Visualizations/.

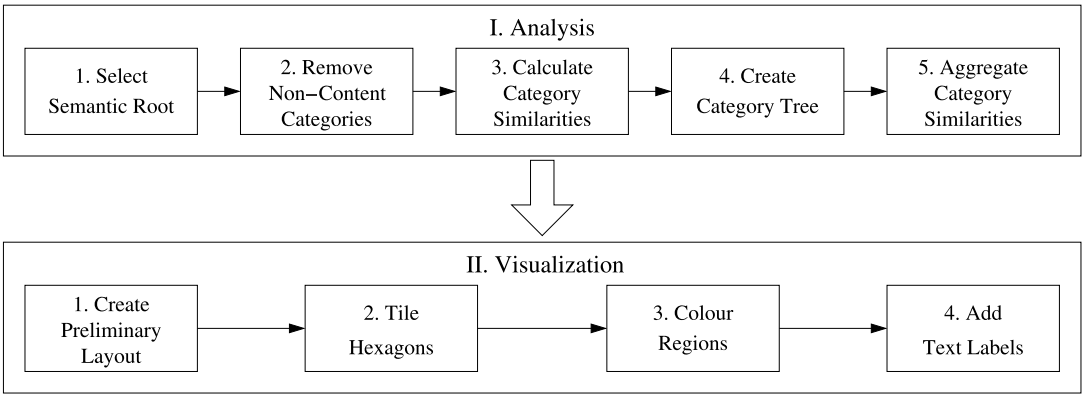


Fig. 2. Map-like analysis and visualization method.

Table 4
Semantic root (shown in bold) in different Wikipedia language editions.

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

Wikipedia administrative categories are placed in a sub-tree of the main Wikipedia category tree. They are usually also identified by a name prefix or special namespace that is unique to each language edition of Wikipedia. In some language editions this sub-tree is not included under the semantic root node, in which case no action is necessary. However, when this sub-tree is included under the semantic root node then we need to remove it. For example, in the Swedish Wikipedia the sub-tree of Wikipedia administrative categories is rooted at category "Wikipedia:Administration", which is placed under the semantic root node "Topp". Pruning the sub-tree rooted at this node removes Wikipedia administrative categories from our data.

Stub categories are those that mark articles "deemed too short to provide encyclopedic coverage of a subject".[4] In many Wikipedia language editions they are contained in the sub-tree of Wikipedia administrative categories. Removing that sub-tree thus also removes stub categories. In those languages in which this is not the case these categories can be found by string matching, as they contain the word "stub" (or its equivalent in the language in question). Thus removal of these categories is possible by looking for this string in the category name.

List categories are those that organize related pages into lists, such as "1910 births" which lists all pages of persons born in 1910. List categories are large in number but not actually useful in giving an indication of the content of Wikipedia. Thus it makes sense to remove these. We rely on the observation that list categories are consistently named (containing common words such as "births", "deaths", "persons", "countries", etc.), and consistently organized (birth and death-related lists are organized into lists by century, which are organized into chronology-related lists). Within a given list, member categories share a common word, which thus helps identify both these list categories and their parent category. We measure the similarity among the names of a pair of sibling categories by doing a character-by-character comparison, counting the number of common characters, and calculating a cosine similarity value (in the range [0, 1]). If the average similarity value of all sibling categories under a common parent category exceeds a pre-defined threshold we determine that these are list categories. We have experimented with different threshold values and found a value of 0.8 to be effective in finding most list categories while avoiding false positives. Table 5 shows an example of list category name matching: parent category "Aircraft 1950–1959" contains the listed sub-categories (only an extract of all sub-categories is shown). By doing pair-wise name comparison and calculating the average matching rate of all pairs, we find that the similarity score for the sub-categories is 0.905, and as this is greater than our threshold 0.8 we conclude that the parent and child categories are list categories that we eliminate from the category tree.

Table 5
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

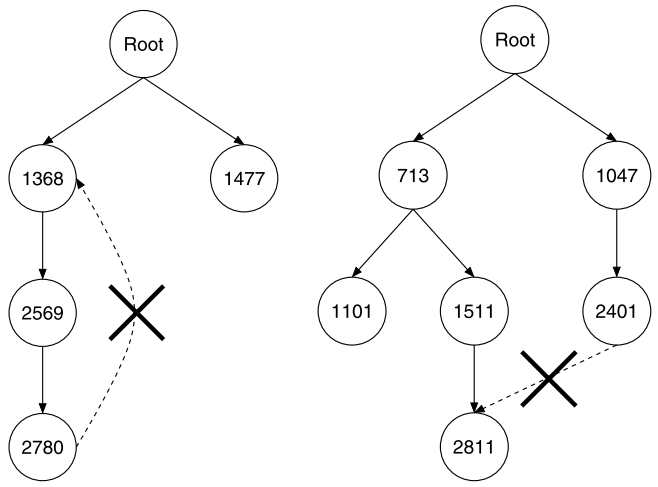
Our removal method eliminates about 5%–10% of categories from the category hierarchy in the different language editions we studied.

Step I.3: *Calculate category similarities*. We use cosine similarity to measure mutual similarity between pairs of categories. Usually 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 termed the *co-occurrence* of category assignments between a pair of categories. A greater number of co-occurrences imply a stronger similarity and vice versa. For a pair of co-occurring categories, their similarity is calculated according to the following formula:

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

where $\cos_{i,j}$ 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.

[4] http://en.wikipedia.org/wiki/Wikipedia:Stub.



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

Fig. 3. Eliminating edges in the category graph (edge indicates "parent category" relationship).

Step I.4: *Create category tree*. The Wikipedia category hierarchy, being the result of manual editing by its users, includes nodes with multiple parents as well as a few loops. To facilitate the analysis process and to simplify our visualization with each category appearing only once, we transform the category graph into a simple directed tree. We first apply a breadth-first search that starts from the chosen semantic root, traversing every node encountered and keeping a list of visited nodes. Loops in the tree are removed by simply eliminating the edge that causes the loop, as illustrated in Fig. 3(a). If the loop involves nodes at different depths in the tree then this approach ensures that each node is connected to the root by the shortest distance; otherwise if the loop occurs at the same depth then no objective decision can be made as to which link to retain (as in [9]), and we retain the link from the older to the younger node considering its creation time. All parent relationships in multiple-parent nodes are removed except for the one to the most similar parent according to the cosine similarity calculated in the previous step, as illustrated in Fig. 3(b).

Step I.5: *Aggregate similarities*. To determine the similarity between a pair of categories it is not sufficient to simply use their direct similarity values, as their similarity is also partly based on the similarities of their sub-categories. In some instances the direct similarity between two categories is zero, because no articles are assigned to both categories, but there may be articles assigned to their sub-categories. In cases like these the articles of their sub-categories should also contribute to a certain degree to the relationship of their parents. This is illustrated in Fig. 4: two categories $c1$ and $c2$ have no co-assigned articles, but their sub-categories $c3$ and $c4$ have co-assigned articles $a1$ and $a2$. In this case the direct similarity of categories $c1$ and $c2$ is zero, but the similarity of sub-categories $c3$ and $c4$ is greater than zero. Intuitively, we would consider categories $c1$ and $c2$ similar because of their child categories' similarity. Thus we define an *aggregated cosine similarity* of a pair of top level categories, which combines both the direct similarity and the similarity from sub-categories. To produce this aggregate we considered different calculation methods, giving more or less weight to the direct similarity, or indeed simply choosing the largest similarity value among all levels of categories.

Given levels of sub-categories $1, \ldots, n$ below a given category, and the similarity of sub-categories of categories C_i and C_j at level k being denoted as $\cos'_{i,j,k}$, we defined the following three formulae:

- Formula A: $\frac{\cos_{i,j} + 2\sum_{k=1}^{n} \cos'_{i,j,k}}{3}$.

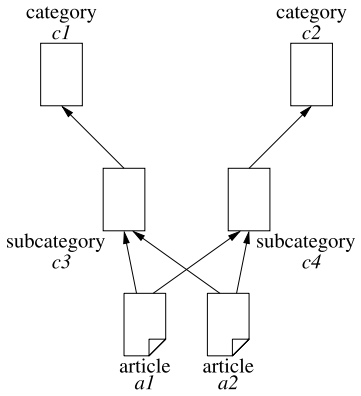


Fig. 4. Articles co-assigned to sub-categories but not to parent categories.

- Formula B: $\frac{\cos_{i,j} + \sum_{k=1}^{n} \cos'_{i,j,k}}{n+1}$.
- Formula C: $\max(\cos_{i,j}, \cos'_{i,j,1}, \ldots, \cos'_{i,j,n})$.

That is, formula A is the weighted sum of direct similarity and aggregated sub-category similarity, with twice the weight on the latter; formula B is the arithmetic mean of direct similarity and all sub-category similarity values; and formula C is the maximum value of all these similarity values. We calculated the aggregate cosine similarity of all unique 276 pairs of the 24 top-level categories in the English Wikipedia and their sub-categories to 5 levels down, using each of the three formulae above. We limited the number of sub-categories to 5 as we found that deeper levels of sub-categories had negligibly small similarity values. A plot of the standard deviations of these similarity values for each pair of top-level categories is shown in Fig. 5. Points on the x axis are the 276 top-level category pairs; the y axis shows the standard deviation of similarity values.

We can observe that the direct similarity alone suffers from the problem of null values where top-level categories do not have any co-assigned articles (this was the case for 242 of the 276 category pairs, i.e. about 88%). Formula C results in greatly varying similarity values with several extreme outliers. Of formulae A and B, formula B (arithmetic mean of direct and sub-category similarities) produced the lowest standard deviation of similarity values. Thus it can minimize the effect of extreme values across categories, while avoiding the null values where direct similarity is zero. For this reason we use this formula for aggregating category similarity values. Once the aggregation is complete, we have a

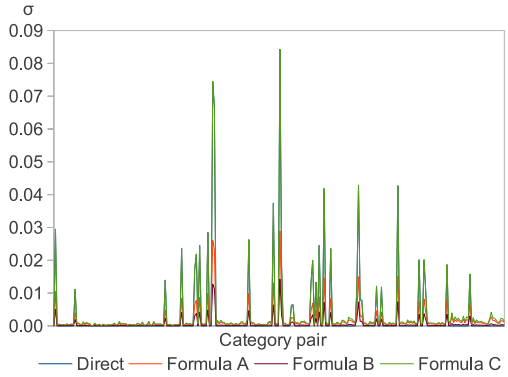


Fig. 5. Comparison of standard deviation σ of direct similarity and similarity aggregation formulae A, B and C for pairs of top-level categories in English Wikipedia.

similarity value for each pair of Wikipedia categories within the chosen n levels of categories below the semantic root.

Step II.1: *Create preliminary layout*. In this step we create a rough layout of the approximate positions and estimated sizes of wiki categories. We use a bottom-up approach to create this layout, proceeding upwards level by level. Based on visual evaluation up to three levels of categories can be well represented without resulting in too many or too small sub-divisions. For example top-level category "Science" contains sub-category "Mathematics", which contains sub-sub-category "Geometry". Following the metaphor of a geographic map, these three levels can be thought of as three levels of political regions namely countries, provinces, 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 [10]. 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. 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 principle of the bottom-up layout is illustrated in Fig. 6: sub-sub-categories are laid out within their related sub-category, as shown in (a); all sub-categories are then laid out in their related top-level category (b). Finally, this entire top-level category (bounded by the outer box) is laid out together with other top-level categories, (c) and (d). As the resulting top-level category areas may overlap each other, we remove any overlaps using the Force Transfer Algorithm of [31]. The result is that all top-level category areas at most touch each other, but do not intersect, as in Fig. 6.

Step II.2: *Tile hexagons*. Each category's region has a certain surface area that is proportional to the size of all of its sub-categories and the number of its articles. The exact shape of this area is (pseudo-)randomly determined by tiling the surface piece by piece starting from the centre point of the area. We use regular hexagons as the basic unit for tiling the surface. Using regular hexagons has two advantages: (1) regular hexagons tile a surface completely, and (2) border lines of areas tiled by hexagons have a natural-looking ragged appearance. We initialize a hexagon lattice in memory. Categories are allocated hexagons in the lattice corresponding to the size and position of categories in the preliminary layout. As the numbers of articles and sub-categories can have extreme variations in some Wikipedia language editions, we apply a logarithmic scale 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 begins at a given starting hexagon, called the region's *pivot point*. We maintain a data structure of already allocated hexagons and grow a region's territory by randomly assigning unused neighbouring hexagons, as illustrated in Fig. 7 (numbers correspond to the order of hexagon selection). Through experimentation we concluded that this form of random assignment produces regions that look most natural, i.e. similar to those in a geographic map. In order to create reproducible output, however, we control randomness by using a fixed random seed. Thus the same input data will always result in the same visualization and make different visualizations comparable with each other.

Step II.3: *Colour regions*. Colour is one of the primary visual elements that a reader perceives in a visualization. Topographic maps often employ colour to represent elevation. In our map we use it to represent a chosen attribute of the Wikipedia data. This could for example be the total number of articles of a category, the average co-author count, a measure of the edit activity in that category, or any other selected attribute the user wishes to visualize. Colouring allows users to quickly spot large and small attribute values and to perceive their distribution throughout the visualization. The colour scheme we employ is similar to that of topographic maps, using a darker colour to represent a higher value.

Step II.4: *Add text labels*. Finally the text labels for category names and selected article titles 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 the text labels to avoid overlaps and produce an optimal appearance. However, given the size of Wikipedia datasets this approach is clearly infeasible. Therefore we place the text labels automatically. Where necessary, we reduce font sizes of the label text to enable labels to remain inside the area of their respective categories, with a minimum allowed font size to ensure readability. Label text styles are used to distinguish category levels. For example top level categories are shown as in "**LIFE**", second level categories as in "BIOLOGY", and third level categories as in "Genetics".

4. Visualizations

We have applied our analysis and visualization method to the English, German, Chinese, Swedish and Danish Wikipedia editions. In all cases we aggregated categories and visualized the first three levels of categories below the semantic root node, using average co-author count for region colouring. The colour scale was the same in all cases, with six colour ranges, each representing a quadruple lower limit relative to the next range (i.e. limits at 1, 4, 16, 64, etc.). Fig. 8 shows the English Wikipedia visualization and extracts from two regions within this visualization. As expected, the differences in average numbers of co-authors per category that were shown in Fig. 1 are clearly evident in the different colouring of category regions. However, it also shows categories of similar numbers of co-authors clustered together. This is illustrated in the two extracts of the map in Fig. 8: the left region (belonging to main category "People") shows sub-categories with medium to high co-author counts, most in the range from 64 to 1024+ average co-authors; on the other hand the region on the right (belonging to main category "Politics") shows that sub-categories there have a much lower co-author count, mostly in the ranges from 1 to 255 average co-authors. Thus there is evidence of patterns of participation in Wikipedia that extend beyond individual categories to clusters of related categories. In other words, popular categories such as "People" include related sub-categories that likewise are popular and thus attract many co-authors who participate in writing articles in these categories. The same appears to be the case for less popular categories in which the

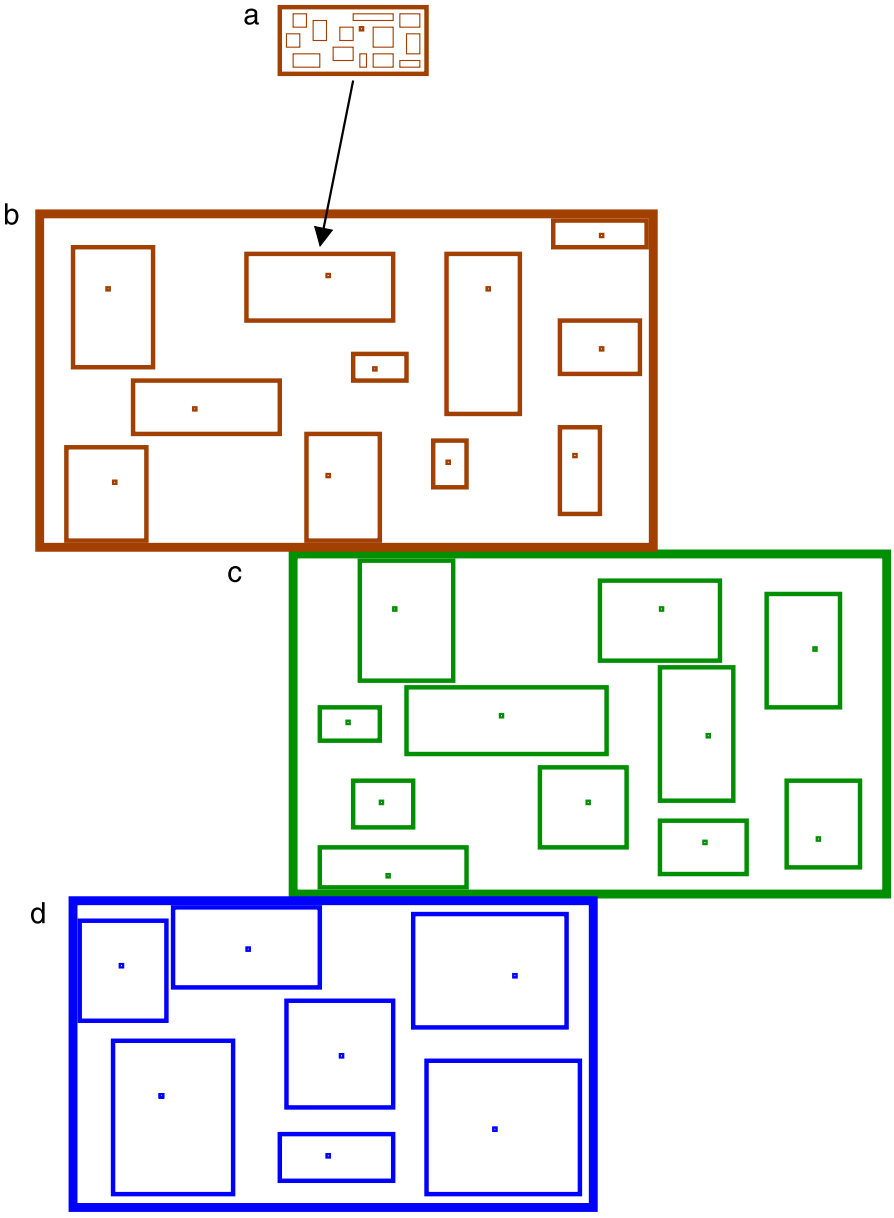


Fig. 6. Preliminary bottom-up layout of categories.

number of co-authors in sub-categories likewise indicates a lower popularity.

Comparing visualizations of different Wikipedia language editions also reveals differences. Fig. 9 shows extracts of the English, German and Chinese Wikipedia maps. Again, the variation of colours corresponding to different average co-author counts is apparent in each of these three maps. However, it is also evident that the colours in the smaller Wikipedias are lighter, indicating a lower co-author count. On the one hand the number of articles and the number of categories are significantly smaller in the smaller wikis. On the other hand, the number of users per article and per category is also smaller (e.g. 4.07 users/article and 23.04 users/category in English Wikipedia vs. 0.95 users/article and 16.88 users/category in German Wikipedia, and a similar situation in the other wikis). As there are fewer human resources to actively edit these smaller wikis, this results in the observed smaller co-author counts. English being an international language spoken far beyond the borders of the countries where it is a native language gives the English Wikipedia a much larger pool of users who are able to contribute to it, and as these numbers show indeed do so. This is not the case for German, Danish and Swedish, none of which enjoy the status of an international language, and thus have a much more limited pool of potential contributors to draw from.

In our visualization similar categories are placed near each other. Performing a visual analysis of entire top-level categories (the categories directly below the semantic root node in the category hierarchy) does not actually reveal any striking similarities in co-authoring. That is, if two similar top-level categories such as, say, "Culture" and "Society" are placed adjacent to each other as they are in the English Wikipedia map, this does not imply that they must have similar co-authoring counts. In fact, looking at entire top-level categories no major differences are apparent. However, looking inside a top-level category, we can visually identify clusters of its sub-categories with the same or similar colouring, i.e. co-author counts. This was already visible in the small extracts of Fig. 8, but it also applies throughout top-level categories—clusters of sub-categories that are adjacent to each other (thus indicating similarity) which have similar average co-author counts.

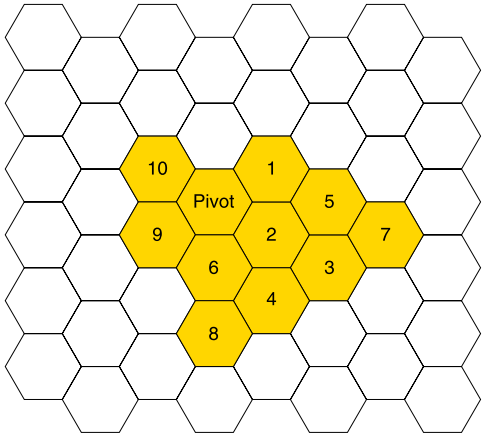


Fig. 7. Example of random hexagon assignment for a category.

An example contrasting different patterns of co-authoring in the English Wikipedia is shown in Fig. 10. Top-level category *Law* is represented in Fig. 10(a). The entire top-level category is dominated by the two colours representing, respectively, 16–63 and 64–255 co-authors. A few sub-categories with 4–15 co-authors exist, and even fewer with 256–1023 co-authors. Thus on the whole there are no extreme differences in the amount of co-authoring that takes place on law-related article content: it appears that the law community pays attention to the development of all law-related topics. Some clusters can be identified, such as the one at centre bottom of the same colour (representing 64–255 co-authors), consisting of sub-category *Law enforcement* and its sub-sub-categories. Relatively fewer co-authors are involved in another two clusters in the upper right: *Legal organizations* and *Legal literature*, mainly with sub-sub-categories in the 4–15 and 16–63 co-author ranges. But the differences of these ranges are not extreme.

A different situation presents itself in Fig. 10(b). Unlike the Law category, the People category shows great extremes of co-authoring: all ranges, from 1–3 co-authors (category *Award winners*) all the way up to 1024 + co-authors (categories *Religious skeptics, Vegetarians, Hindus, Christians* and *People self-identifying as substance abusers*) are represented. Among different sub-categories there are obvious differences in authoring interest that these attract: a string of category clusters in the left half of the figure includes many categories with high co-author counts. From left to right these are related to people with disorders, people by ideology, and religious followers—the first of which commands great personal interest by the concerned people and their families, and the latter two tend to be hotly debated topics that consequently attract much attention. However, this is bordered by a cluster of categories in the centre belonging to the *Biography* sub-category, and on the far right in the sub-category *Human names*, both of which have only low to medium co-author counts (mostly ranging from 16 to 254). We can only surmise that these extreme differences between and within sub-categories in the People category are due to the more informal content of this category, which attracts widely differing attention and popularity, compared with more professional-oriented content such as in the Law category which may be more carefully curated by members of the legal profession.

Trying to perceive these patterns of distribution by only looking at Wikipedia article pages would be practically nearly impossible. The visualization, however, by representing the multi-dimensional mutual relations of categories through proximities and the co-authoring attribute through colour, makes these patterns "pop out" of the visualization and be immediately perceived.

5. Evaluation

In order to assess our visualization we conducted a usability evaluation, focusing mainly on usability, accuracy, speed, and preference. As no other visualizations of co-authorship exist, we compared our visualization against textual tables of summary data on co-authorship that we extracted from Wikipedia. This data consisted of the first three levels of categories in the Simple English Wikipedia together with an average co-author count for each (a 20-page PDF document) and a table of similarity values for pairs of categories in the first three levels of categories (a 47-page PDF document). The visualization was also of the Simple English Wikipedia of the same date (October 2012) and was presented as a PNG file in an image viewer application. We set up an experiment with two groups of subjects, each performing two rounds of evaluation. In round 1 group 1 performed a set of tasks with the help of our visualization, and in round 2 a similar set of tasks with the help of the textual data only. Group 2 then performed the same two rounds of evaluation, but with a reversed order of the information tool: first with the textual data only, and then with the visualization. This within-subject design allowed us to control for the learning effect that occurs when the same kinds of tasks are performed twice.

We recruited 28 subjects for this evaluation, divided in two groups of 14 subjects each. All were students of our university: 1 Bachelor student, 23 Master students and 4 Ph.D. students. There were 12 males and 16 females, aged 21–33 with mean and median age of 24. Of these, 17 students had a technical major (computer science and engineering), and the other 11 a non-technical major (business, accounting, and Chinese). Self-rating their IT skills, 18 responded they had basic skills, 2 had advanced skills, and 8 had programming skills. Finally, when asked about their prior Wikipedia experience 26 responded they had read Wikipedia articles, 1 had edited Wikipedia articles, and 1 had never used Wikipedia before.

We gave our subjects the following set of tasks (these are the tasks for round 1; in round 2 the tasks were very similar):

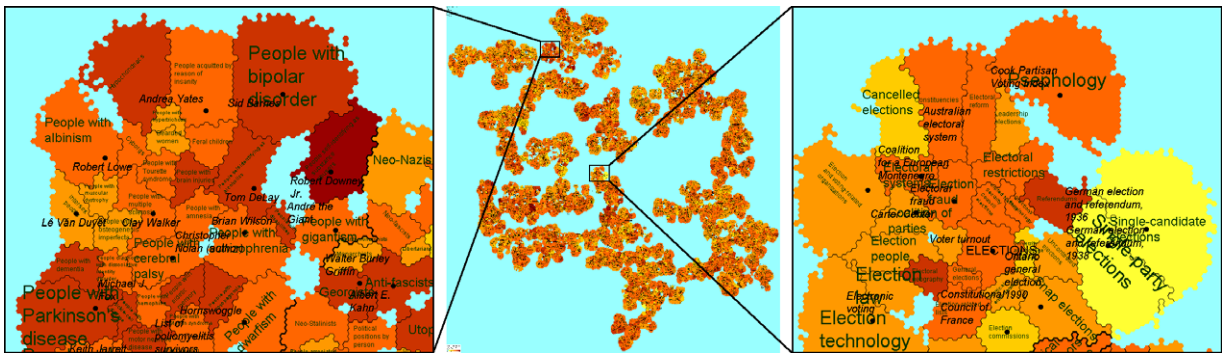


Fig. 8. English Wikipedia map (centre) and extract of categories with medium to high (left) and low to medium (right) average co-author counts.

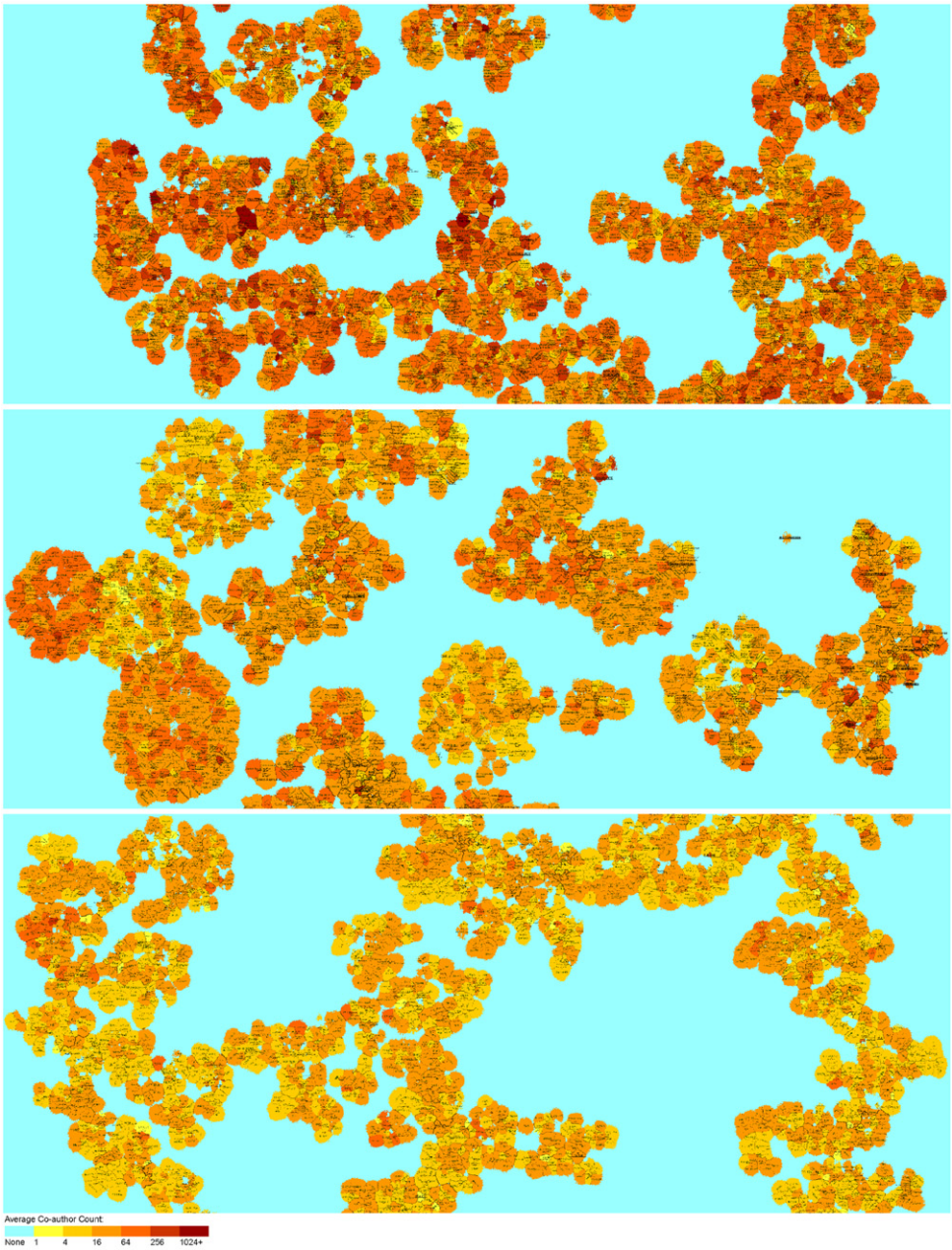


Fig. 9. Extract of Wikipedia maps of English (top), German (middle) and Chinese (bottom) Wikipedias, with the same colour coding of average co-author counts.

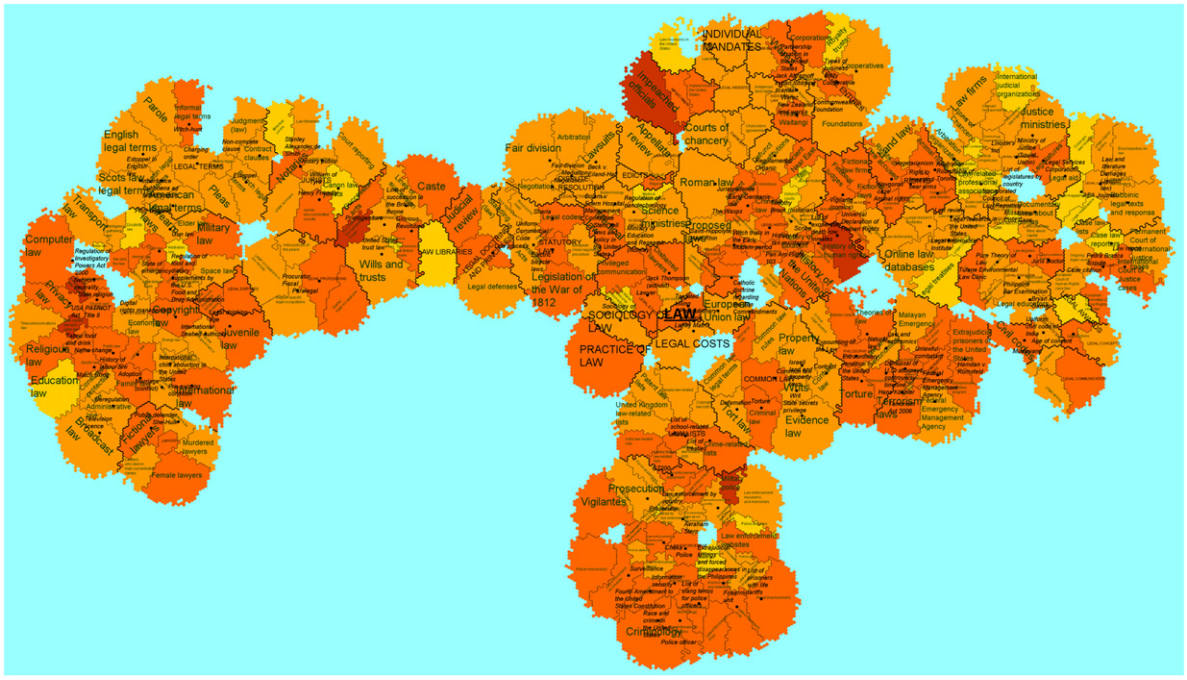
1. Name the level-1 category that has the largest number of sub-categories.
2. Name the most active level-1 category (meaning that it has many sub-categories with a high co-author count, $\geq$64).
3. Name the category that has the highest co-author count.
4. Name the category within the level-1 category *Science* that has the highest co-author count.
5. Name the level-1 categories that are most closely related to level-1 category *Religion*.
6. Within level-1 category *Literature* there is level-2 category *Fiction*. Name three other level-2 categories that are most closely related to category *Fiction*.

These six tasks focused on different kinds of information: (i) overviews of level-1 categories (tasks 1 and 2); (ii) details of specific categories (tasks 3 and 4); and (iii) relationships between categories (tasks 5 and 6).

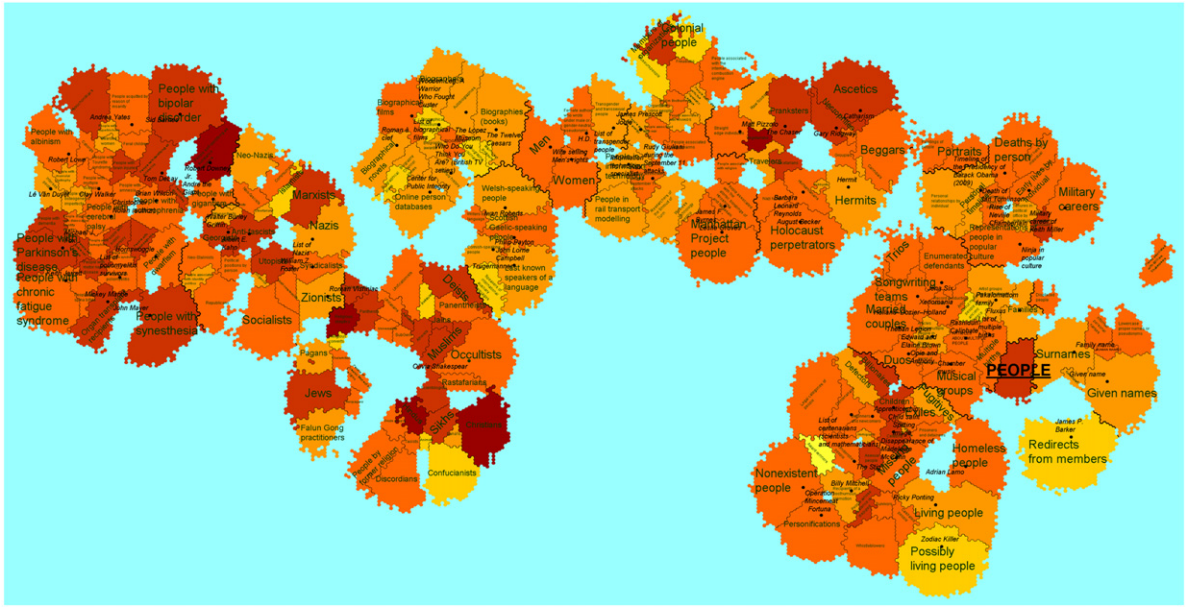
Each round of tasks was followed by an exit questionnaire asking subjects to assess usability of the information tool just used (visualization vs. tables of category data). Subjects gave feedback using a 7-point Likert scale (from strongly disagree through neutral to strongly agree) on the following five statements (for the category data tables, the statements correspondingly referred to the category data):

1. The visualization was easy to learn.
2. The visualization was easy to use.
3. It was easy to understand the information presented by the visualization.
4. It was easy to perform the evaluation tasks by using the visualization.
5. I was satisfied with using the visualization for performing the evaluation tasks.

We also assessed accuracy of the answers given and measured the speed of task performance from start to finish of each evaluation round. Finally, to assess preference, at the end of the evaluation subjects were asked to respond to the statement (again



(a) Top-level category *Law*.



(b) Top-level category *People*.

Fig. 10. Samples of top-level categories in English Wikipedia.

with answers on the same 7-point Likert scale as before): "I prefer using the visualization instead of the category data".

The tasks we gave our subjects were challenging, particularly using the textual category data and the limited time (15 min in each round). Even with our visualization it was not easy to perform all tasks correctly. However, we observed a greatly higher accuracy using the visualization than the category data, as shown in Table 6: using the category data only about 1 in 5 answers was correct, whereas with the visualization half of all answers were correct—a 2.6 times higher accuracy.

Speed of task performance was also better when using our visualization, as shown in Table 7. On average, tasks were completed 2.5 min faster with our visualization—using the textual category data our subjects required about 25% more time.

Table 6
Mean accuracy (0..1).

	Visualization	Category data	Mean
Group 1	0.45	0.25	0.35
Group 2	0.54	0.12	0.33
Mean	0.50	0.19	0.34

Comparing the accuracy and speed between groups we can observe that whenever an information tool was used in round 2 it allowed faster task completion than when the same information tool was used in round 1 (recall that Group 1 used the visualization in round 1 and the category data in round 2, whereas Group 2 used the category data in round 1 and the visualization in round 2).

Table 7
Mean speed of task performance (mm:ss).

	Visualization	Category data	Mean
Group 1	10:51	11:47	11:19
Group 2	09:26	13:30	11:28
Mean	10:09	12:39	11:24

Table 8
Usability results.

Factor	Visualization		Category data		Significance
	Mean	StdDev	Mean	StdDev	*p*
Ease of learning	5.50	1.04	4.54	1.77	0.0136*
Ease of use	5.21	1.26	3.93	1.63	0.0005***
Ease of understanding	5.57	1.20	4.82	1.70	0.0321*
Ease of performing tasks	5.04	1.45	3.25	1.58	0.0001***
Satisfaction	4.96	1.50	3.43	1.69	0.0002***

Comparing with the mean value of both rounds combined, the improvement over the mean accuracy was 8% for the visualization and 32% for the category data. The speed-up over the mean speed was about 8% for the visualization, and about 7% for the category data. We conclude that a learning effect helped improve accuracy and speed of task performance.

Usability results indicated a higher usability of the visualization in all aspects measured; see Table 8. The difference was most pronounced for the ease with which tasks could be performed, followed by satisfaction. We tested for statistical significance using a two-tailed paired t-test and found the difference to be statistically significant: at the 5% level for ease of learning and ease of understanding, and at the 0.1% level for the other three usability factors.

Finally, the feedback on the preference statement indicated general agreement, with more than half of the subjects answering that they agreed (11 subjects) or strongly agreed (5 subjects), with mean of 5.0, median 6, mode 6, and standard deviation 1.8. That is, overall our subjects were of the opinion that the visualization was preferable to the textual category data.

In summary, the evaluation established that our visualization has higher usability, facilitates faster task performance, supports higher accuracy, and is preferred over the textual category data alternative.

6. Applications

In this section we briefly discuss other potential applications of our visualization. In the previous work we have used our visualization method to represent a different attribute, namely article count [32]. Region colouring in that visualization indicated how large a given category is in terms of the number of articles. Visualization of other attributes is likewise possible, such as number of revisions, article age, recent edit activity, etc. Doing so can be of use to various stakeholders: site administrators, content curators, editors, authors, managers, and researchers, to name a few obvious ones. As wikis find more widespread application within organizations, and they grow in volume and use, more and more stakeholders will want to know how their wikis evolve, whether they continue to grow, whether all their content is well developed, and so on. Our visualization method facilitates such analyses.

6.1. Understanding a wiki's category composition

Wiki content is typically organized in categories, and to understand the current topic distribution in a wiki we can look at the distribution of articles over the existing categories. Our map-like visualization provides an easily perceivable overview of the category composition of a given wiki. As area sizes relate to the number of articles of the corresponding category, relative sizes can easily be perceived which may reflect differences in popularity of topic areas among authors. Such sizes may also indicate the levels of content maturity, or alternatively a need for further content development. Editors can quickly perceive the current content collection of a wiki, identifying large topic areas for potential division into sub-categories, or for identifying relatively under-represented topics that require more attention.

6.2. 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. When viewing the visualization, proximities clearly express the relationships between categories. Examples of this could be seen in Fig. 8 where in the right excerpt a number of closely related sub-categories of "Politics" were placed adjacent to one another ("Elections", "Election people", "Election technology", and other election-related categories). Thus the visualization summarizes the overall relationship of these categories into visual proximity. Given the multi-dimensional relationships between categories, our form of visualization provides a straightforward way to discover these relations among topic areas in a wiki that other more traditional forms of representation such as hierarchical folder structures, which show only parent–child relationships but not relationship strengths, cannot show.

6.3. Overview of multiple wikis

A map-like visualization effectively serves as a "world map" of a wiki. It provides the 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 wikis makes comparisons possible, just as political maps help compare aspects of different physical countries. The first impression we get from such exploration of the visualizations may indicate differences of content distribution, maturity, popularity, etc.

6.4. Comparison of topic areas across 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. For example, we compared the "Science" top-level category in the Danish, Swedish and Chinese Wikipedias and found that it occupies the largest area in the Swedish Wikipedia, followed by the Chinese and Danish ones. Also, many areas in the Danish and Swedish visualizations are displayed in darker colours, indicating a higher density of articles (as in our previous work colour represented the number of articles) whereas most Chinese sub-category regions were displayed in lighter colours. These conclusions agree with the statistical data, but are easier to perceive when the entire map is viewed in overview, compared with reading long lists of text and numbers.

7. Conclusions

Many of today's internet-scale applications produce large amounts of data, and this is expected to increase even more in the years to come. These "big data" applications produce terabytes, exabytes and more of data [33]. Making effective use of this data becomes increasingly difficult because of its sheer volume. Even a single website such as Wikipedia stores terabytes of data. In this article we presented a novel method for analyzing large amounts of wiki data and visualizing it in a form similar to a geographic map. We applied our method to data from five Wikipedia language editions: English, German, Chinese, Swedish and Danish, and have discussed further potential applications of our visualizations. Visualizations such as these have the potential to reveal in a readily perceivable way much information contained within large wiki article collections that is otherwise difficult to perceive.

The contributions of our work are two-fold: firstly we propose using a graphical map-like representation to allow users to visually perceive relations between wiki categories through proximity and colouring; secondly we have devised a method for transforming wiki data into map form which has the benefit of being more easily understandable by a wide range of users compared with other more specialized visualizations.

Our research is 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. Currently processing English Wikipedia data on a single server machine takes many hours and even days. A move to grid/cloud computing such as the one described in [34] would significantly aid the timely processing of data and generation of visualizations. Finally, we are planning to include more map elements to represent other aspects of the wiki data, such as road networks representing significant linkages between articles. This will enrich our visualization by increasing the information it communicates.

References

[1] D. O'Leary, Wikis: 'from each according to his knowledge', Computer 41 (2008) 34–41.
[2] F.B. Viégas, M. Wattenberg, K. Dave, Studying cooperation and conflict between authors with history flow visualizations, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2004, pp. 575–582.
[3] B. Suh, E.H. Chi, A. Kittur, B.A. Pendleton, Lifting the veil: improving accountability and social transparency in Wikipedia with WikiDashboard, in: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, ACM, 2008, pp. 1037–1040.
[4] T. Holloway, M. Bozicevic, K. Börner, Analyzing and visualizing the semantic coverage of Wikipedia and its authors, Complexity 12 (3) (2006) 30–40.
[5] F. Ortega, J.M. Gonzalez Barahona, Quantitative analysis of the Wikipedia community of users, in: Proceedings of the 2007 International Symposium on Wikis, ACM, 2007, pp. 75–86.
[6] P.K.-F. Fong, R.P. Biuk-Aghai, What did they do? Deriving high-level edit histories in Wikis, in: Proceedings of the 6th International Symposium on Wikis and Open Collaboration, ACM, 2010, pp. 2:1–2:10.
[7] A.G. West, I. Lee, What Wikipedia deletes: characterizing dangerous collaborative content, in: Proceedings of the 7th International Symposium on Wikis and Open Collaboration, ACM, 2011, pp. 25–28.
[8] J. Yu, J.A. Thom, A. Tam, Ontology evaluation using Wikipedia categories for browsing, in: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, ACM, 2007, pp. 223–232.
[9] T. Zesch, I. Gurevych, Analysis of the Wikipedia category graph for NLP applications, in: Proceedings of the TextGraphs-2 Workshop, NAACL-HLT, NAACL, 2007, pp. 1–8.
[10] T. Kamada, S. Kawai, An algorithm for drawing general undirected graphs, Information Processing Letters 31 (1989) 7–15.
[11] I.S. Dhillon, D.S. Modha, Concept decompositions for large sparse text data using clustering, Machine Learning 42 (2001) 143–175.
[12] Y.H. Li, A.K. Jain, Classification of text documents, The Computer Journal 41 (1998) 537–546.
[13] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., 1986.
[14] J. Szymański, Mining relations between Wikipedia categories, in: Networked Digital Technologies, Part II, Springer, 2010, pp. 248–255.
[15] M. Wattenberg, F.B. Viégas, 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, Springer, 2007, pp. 272–287.
[16] M. Blades, J.M. Blaut, Z. Darvizeh, S. Elguea, S. Sowden, D. Soni, C. Spencer, D. Stea, R. Surajpaul, D. Uttal, A cross-cultural study of young children's mapping abilities, Transactions of the Institute of British Geographers 23 (1998) 269–277.
[17] A. Skupin, The world of geography: visualizing a knowledge domain with cartographic means, Proceedings of the National Academy of Sciences 101 (Suppl. 1) (2004) 5274–5278. National Academy of Sciences.
[18] T. Kohonen, Self-Organizing Maps, Springer, 2001.
[19] I. Jolliffe, Principal Component Analysis, second ed., Springer, 2002.
[20] J. Yang, M.O. Ward, E.A. Rundensteiner, S. Huang, Visual hierarchical dimension reduction for exploration of high dimensional datasets, in: Eurographics/IEEE TCVG Symposium on Visualization, Eurographics Association, 2003, pp. 19–28.
[21] S. dos Santos, K. Brodlie, Gaining understanding of multivariate and multidimensional data through visualization, Computers & Graphics 28 (2004) 311–325.
[22] W. Müller, T. Nocke, H. Schumann, Enhancing the visualization process with principal component analysis to support the exploration of trends, in: Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation—Volume 60, Australian Computer Society, 2006, pp. 121–130.
[23] M.J. McQuaid, T.-H. Ong, H. Chen, J.F. Nunamaker Jr., Multidimensional scaling for group memory visualization, Decision Support Systems 27 (1999) 163–176.
[24] M. Williams, T. Munzner, Steerable, progressive multidimensional scaling, in: Proceedings of the IEEE Symposium on Information Visualization, IEEE Computer Society, 2004, pp. 57–64.
[25] A. Buja, D.F. Swayne, M.L. Littman, N. Dean, H. Hofmann, L. Chen, Data visualization with multidimensional scaling, Journal of Computational and Graphical Statistics 17 (2008) 444–472.
[26] P. Sorg, P. Cimiano, Cross-lingual information retrieval with explicit semantic analysis, in: Working Notes for the CLEF 2008 Workshop.
[27] M. Potthast, B. Stein, M. Anderka, A Wikipedia-based multilingual retrieval model, in: C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, R. White (Eds.), Advances in Information Retrieval, in: Lecture Notes in Computer Science, vol. 4956, Springer, 2008, pp. 522–530.
[28] P. Sorg, P. Cimiano, Enriching the crosslingual link structure of Wikipedia—a classification-based approach, in: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence, WikiAI'08.
[29] A. Hautasaari, T. Takasaki, T. Nakaguchi, J. Koyama, Y. Murakami, T. Ishida, Multi-language discussion platform for Wikipedia translation, in: T. Ishida (Ed.), The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability, Springer, 2011, pp. 231–244.
[30] V. Nastase, M. Strube, B. Börschinger, C. Zirn, A. Elghafari, Wikinet: a very large scale multi-lingual concept network, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10, European Language Resources Association (ELRA), 2010.
[31] X. Huang, W. Lai, Force-transfer: a new approach to removing overlapping nodes in graph layout, in: The Twenty-Fifth Australasian Computer Science Conference, ACSC2003, pp. 349–358.
[32] C.-I. Pang, R.P. Biuk-Aghai, Wikipedia world map: method and application of map-like wiki visualization, in: Proceedings of the 7th International Symposium on Wikis and Open Collaboration, ACM, 2011, pp. 124–133.
[33] L.M. Surhone, M.T. Tennoe, S.F. Henssonow, Big Data, Betascript Publishing, 2010.
[34] R.L. Grossman, Y. Gu, M. Sabala, W. Zhang, Compute and storage clouds using wide area high performance networks, Future Generation Computer Systems 25 (2009) 179–183.

Robert P. Biuk-Aghai is an Assistant Professor of Computing Sciences at the University of Macau. He holds a Ph.D. degree in Computing Sciences from the University of Technology, Sydney, and an M.Sc. degree in Information Systems from the London School of Economics. Dr. Biuk-Aghai's research interests include collaboration systems, information visualization, and mobile GIS.

Cheong-Iao Pang is a Ph.D. candidate in the Department of Computing and Information Systems of the University of Melbourne. He obtained an M.Sc. degree from the University of Macau. He is interested in understanding why people look for information, and how to improve this search process with information visualization and interactive software.

Yain-Whar Si is an Assistant Professor at the University of Macau. He holds a Ph.D. degree in Information Technology from the Queensland University of Technology, Brisbane, and an M.Sc. degree in Software Engineering from the University of Macau. His research interests are in the areas of business process management and decision support systems.