Visualizing Co-Authorship Networks in Online Wikipedia

Robert P. Biuk-Aghai
Department of Computer and Information Science
Faculty of Science and Technology
University of Macau, Macau S.A.R., China
Tel: +853-3974365, Fax: +853-28838314
Email: robertb@umac.mo

Abstract—The Wikipedia online user-contributed encyclopedia has rapidly become a highly popular and widely used online reference source. However, perceiving the complex relationships in the network of articles and other entities in Wikipedia is far from easy. We introduce the notion of using co-authorship of articles to determine relationship between articles, and present the WikiVis information visualization system which visualizes this and other types of relationships in the Wikipedia database in 3D graph form. A 3D star layout and a 3D nested cone tree layout are presented for displaying relationships between entities and between categories, respectively. A novel 3D pinboard layout is presented for displaying search results.

I. INTRODUCTION

The Wikipedia online encyclopedia [1] has over the past few years become a highly popular and widely used online reference work on the Internet. It is a web-based free-content encyclopedia, meaning that its users have the right to edit its articles. As of September 2006 it had over five million articles in over 200 languages, of which close to 1.4 million articles (27%) were in English. Because of the very large number of articles, which typically are linked to each other in complex ways, it can be difficult to perceive the relationships an article has with other articles and entities in the Wikipedia database. For example, it may be difficult to perceive whether an article is related to few or many other articles, how closely its content is related to that of other articles, and how popular the different articles in a collection of related articles are. For the viewer of a particular article, it may be difficult to know which of the (possibly many) other articles that this article links to are most closely related to it.

To answer these kinds of questions, we have developed an information visualization system called WikiVis (short for Wikipedia Visualization). This system produces various different 3D graphical representations of the content of a Wikipedia database. It supports the exploratory navigation of the large Wikipedia space, and also provides a search function with graphical display of its results. Several novel visualization layout schemes were developed for this system. The work presented is of significance beyond the Wikipedia system, to any kind of collaborative authoring system, including wiki systems in general, blogs, and others.

The remainder of this paper is organized as follows: Section II introduces the structure of the Wikipedia database and the entities represented in it. Section III presents our WikiVis system, and Section IV discusses the layout methods developed for it. Finally, Section V concludes this paper.

II. WIKIPEDIA DATA

The Wikipedia system maintains 18 different types of entities, including articles, categories, images, article discussions, etc. The most important of these is the article entity which corresponds to an article in the online encyclopedia. Another important type of entity is the category; it corresponds to a defined category of articles under which a number of actual article instances may be found. A given article may be labeled with zero or more categories. For example, on 2006/08/01, the Wikipedia article entitled “Scientific visualization” was labeled as belonging to the five categories “Computational science”, “Computer graphics”, “Visualization”, “Scientific modeling”, and “User interface techniques”.

Wikipedia entities are stored in a relational database. Each type of entity is described in the database by a number of attributes, including its title, text, last editing user, last edit timestamp, etc. Moreover, the entire revision history of articles and several other types of entities is preserved in the Wikipedia database. This makes it possible to review every revision of an article. The operators of Wikipedia make regular dumps of the Wikipedia database available for free download [2]. The database dumps tend to be very large—for instance, the download files of the 2006/08/16 dump of the English language edition of Wikipedia, including all revisions of all pages, compressed with SevenZip, measured about 5.1GB. The uncompressed version occupies about 100 times that space. For our development and testing purposes we have been using the much smaller Chinese language edition, which currently measures nearly 260MB for the SevenZip-compressed file of all page revisions, with a total of nearly 90,000 articles.

Wikipedia articles consist of hypertext, i.e. they make extensive use of hyperlinks to other locations within the same article, to another article, or even to web pages outside Wikipedia. Hyperlinks are rendered in the standard way, usually as underlined and differently-coloured text. This makes it possible to identify other content that is (usually) related to that of the current page.
However, not all articles that are related are linked in this way, as links need to be manually created and the linking is thus subject to the authors performing this task. In order to find articles that are related to a given article, an alternative approach is to find those articles that have a significant degree of co-authorship with a given article. Co-authorship of two articles means that the intersection of their sets of authors is non-empty. The degree of co-authorship is a measure of the dimension of this co-authorship relative to the size of the set of authors of a given article. Let the set of authors (contributors) of each article \( a \) be defined as \( C_a = \{c_1, c_2, \ldots, c_n\} \) with \( n \) being the number of distinct authors. Then the degree of co-authorship of a pair of articles \( \{a_1, a_2\} \) is defined as:

\[
\Phi_{(a_1,a_2)} = \frac{|C_{a_1} \cap C_{a_2}|}{|C_{a_1}|}
\]

For illustration: given an article \( a_1 \) with a set of contributors (authors) \( C_{a_1} = \{c_1, c_2, c_3, c_4, c_5\} \), and a different article \( a_2 \) with a set of contributors \( C_{a_2} = \{c_1, c_2, c_3, c_6\} \), the degree of co-authorship \( \Phi_{(a_1,a_2)} = 3/5 = 0.6 \). A larger value for the degree of co-authorship indicates that a given pair of articles have more authors in common.

Co-authorship is significant as it usually co-incides with co-expertise. A high degree of co-authorship is thus a good predictor for content similarity. This relies on the observation that the same group of authors are highly likely to contribute to multiple articles of similar content, while they are highly unlikely to contribute to articles of very different content. For example, a given group of authors who contributes to several articles related to computer graphics and visualization is not likely to contribute to articles about chimpanzees or medieval English literature. Note that it is possible that an individual author with expertise in several different areas contributes to widely different articles, but this is for a single author, not for an entire group of authors co-contributing to these different articles.

Co-authorship is thus an important property of pages in Wikipedia as well as other collaborative authoring systems for identifying related content. The following section introduces the system we have built for visualizing, among other things, the degree of co-authorship in Wikipedia.

### III. WikiVis

In order to identify the co-authorship relationships a given article has with a set of other articles, we visualize the set of relationships as a three-dimensional (3D) graph. This visualization has the main purpose to support the exploration of the large Wikipedia space through the co-authorship relationship.

The choice of a three-dimensional visualization was motivated by the desire to limit the total required screen space of the visualization. The number of objects that can appear in a single visualization may number in the hundreds, which could require substantial amounts of screen space for display at a readable size if layered out in only two dimensions. Using three dimensions, a larger number of objects can be accommodated within a given screen area. As usual with three-dimensional layout, there is a price to be paid for this more efficient screen area utilization as objects facing the viewer may obscure other objects located behind them. This, however, is compensated for by means for manipulating the 3D visualization, including rotation, zooming, and translation, all of which we have implemented in our WikiVis system.

#### A. Entity View

The main type of visualization is the entity view. It shows co-authorship relationships between a Wikipedia entity and all other related Wikipedia entities, regardless of type. Typically it is used to show other Wikipedia entities related to a given Wikipedia article.

An article in Wikipedia may contain many links to other articles, images, discussions, categories, help pages, etc. These are visible as hyperlinks in the article page itself. An example is shown in Figure 1 (a) for the main page of the Chinese language edition of Wikipedia. A corresponding entity view of this page is shown in Figure 1 (b).

It displays the main page as the node at the centre of the graph. Other nodes surrounding the centre node represent other entities in the Wikipedia database that are related to it. Different shapes and colours represent different types of entities. In the entity view of Figure 1, the majority of the other nodes represent articles (shown as light blue cubes). There are also several images (red cones), categories (yellow cubes), and other types of entities. Distance of a node from the central node is inversely proportional to the degree of co-authorship between that pair of nodes.

An entity view complements the display of a Wikipedia article in the web browser by providing a display of that article’s context. The article itself can be considered the focus of display. Combined, the article and the entity view thus provide the user a focus + context view.

Moreover, a given node in the entity view may be expanded to display a visualization of the nodes related to that node. This can then be repeated on the newly-expanded nodes, etc. An example entity view with several expanded nodes is shown in Figure 2 (a). The newly expanded node is shifted to the centre of the display, with the other expanded nodes remaining in their expanded state. This contributes to a focus + context view within the entity view: the article node of current interest is easily identified, while the recent navigation history remains visible for context.

The display of node labels may be switched between a pop-up display mode which displays the label of one node only (the one which the mouse hovers over, as shown in Figures 1 and 2 (a)), and a display of node labels for all visible nodes, as is shown in Figure 2 (b).

#### B. Category View

Besides the display of co-authorship relationships through entity views, a second form of visualization provided by WikiVis is the category view which displays the hierarchy of categories in Wikipedia, as well as the articles within a given category. The collection of categories in Wikipedia is
user-editable and thus extensible. Currently it numbers several thousand. All categories are organized in a strictly hierarchical fashion, with super- and sub-categories, and with the single root category "Categories". An example of a category view containing several levels of categories is shown in Figure 3. Categories are represented by yellow cubes. Articles belonging to a certain category are also shown (displayed as light blue cubes), and other types of entities belonging to a certain category may also appear (displayed by different shapes and colours). In the category view of Figure 3 several article nodes can be seen on the bottom level of the category hierarchy, as well as a few article and other nodes elsewhere.

As with the entity view, the category view can be navigated by expanding a given category node, and this can be repeated until the lowest level at which only non-category nodes appear. Moreover, the display of node labels too can be switched between one-node display on mouse-over and all-node display.

C. Search Result View

The final form of visualization in WikiVis is the search result view. WikiVis includes a search function which searches the Wikipedia database for all entities that match a user-specified search term. The result of a search is displayed in a search result view. An example of this is shown in Figure 4.

The search result view displays a node representing the search term at the centre of a plane, and the matching entities
Numerous graph layout algorithms have been developed for information visualization (see [3] for an excellent survey). Depending on different factors, such as the total number of nodes in the graph, the degree of connectivity between nodes, the need for concurrent display of all or some of the nodes, and others, various algorithms are preferable to others. For visualizing relationships between Wikipedia entities, the first observation is that the number of nodes in a graph may be very large. Even when visualizing only the entities related to a single article the nodes may number one hundred or more. When navigating through neighbouring nodes, the total number of nodes quickly increases to hundreds, or even thousands. If details of individual nodes are to remain visible, two-dimensional graph layout is usually not adequate for visualizing this large number of nodes on the screen without overlapping. Therefore we have decided to use three-dimensional graph layout in our WikiVis system. Although this may result in some nodes being obscured by other nodes positioned on top of them, three-dimensional graph manipulation, such as rotating, makes it possible to adjust the view of the graph so as to bring these nodes to the front and thus make them visible. The result is a compact layout of a large number of nodes with details such as node labels visible for at least hundreds of nodes.

### IV. Layout Methods

The entity view, displaying the relationships between entities in the Wikipedia database, shows a three-dimensional graph with a large number of nodes surrounding a single central node. This layout method is similar to the well-known radial layout. To accommodate a larger number of nodes at each level, we have modified the radial layout slightly to arrange child nodes in a semi-spherical area around the parent node. This modified graph layout is called a star layout, as the edges connecting related nodes to the central node emanate from the centre similar to the rays of light emanating from a star—see Figure 1 for an example.

In our entity view, the edge length is inversely proportional to the degree of co-authorship between nodes. The degree of co-authorship $\Phi$, defined in Section II above, is a value in the interval $[0,1]$: when the intersection of the set of editors is the empty set, the value is 0 meaning the articles have no co-authorship relationship; on the other hand, when the intersection of the set of editors is equal to the size of the article under comparison, the value is 1 and the articles have the strongest relationship. A threshold $t$ is established, above which the relationship between two articles is considered significant. This is to prevent all articles to display relationships to all other articles, which would make the visualization unreadable and of little value.

For pairs of articles $m$ and $n$ for which $\Phi_{m,n} > t$, the length $l$ of the edge between the two corresponding nodes is calculated as the inverse proportional of the edge length constant $k$:

$$l_{m,n} = \frac{1}{0.1 + \Phi_{m,n}} \times k$$

The resulting length $l_{m,n}$ is a value in the interval $[\frac{10k}{11}, 10k]$. That is, the most unrelated article nodes are eleven times as distant from the central article node as compared to the most related article nodes. The constant $k$ itself is empirically determined and can be adjusted as necessary to improve the visual quality of the display.

For an initial star layout, i.e. when only one star is displayed, the positions of nodes around the central node are calculated so as to evenly distribute nodes in the star. However, for subsequent stars, i.e. expansions of nodes in the initial star, the positions of nodes are constrained to a semi-sphere whose bottom plane is perpendicular to the edge connecting its central node to the node of the star it is related to. This is to avoid an overlap between nodes in neighbouring stars. The effect is that the layout of stars other than the initial one somewhat resembles that of the radial view [4]—see Figure 2 (a) for an example.
B. Nested Cone Tree Layout

The category view, displaying relationships between Wikipedia categories, and between categories and articles, shows a three-dimensional graph with a large number of nodes connected in a strictly hierarchical fashion. The layout method used is a variation of the well-known cone tree layout method [5]. We term our variation the nested cone tree layout—see Figure 3 for an example.

This layout is much simpler than the star layout discussed above. The only type of relationship visualized here is the parent-child relationship that exists between a category and its sub-categories, or between a category and its articles. Because this is a strictly hierarchical relationship, the application of the cone tree layout method would appear suitable. However, the original cone tree layout only lays its nodes out in a single circle along the rim of the cone. Given the very large number of Wikipedia categories, numbering in the thousands, application of the original cone tree layout method would result in very wide-angled cones with large rims, each with a circle containing hundreds or even thousands of nodes. Moreover, at the centre of these cones there would be a large open space. Thus the space-utilization of the original cone tree layout is not optimal for large numbers of nodes.

We have modified the original cone tree layout method by introducing the notion of multiple nested cones. Each cone is of equal height but with a different angle (and thus a different rim radius). The tips of all the cones coincide and the cones point in the same direction. As a result, the cone rims along which nodes are placed form multiple concentric circles. An example of this is shown in Figure 5 which contains about 1,100 nodes arranged in 14 concentric circles. The circles are spaced at slightly increasing distances from the centre out, with similar spacing between nodes in each circle. If the original cone tree layout were used for the same number of nodes, a single very large circle with a large empty space at the centre would have resulted.

C. Pinboard Layout

The search result view, displaying the result of applying a search to the Wikipedia database, shows a three-dimensional graph consisting of nodes placed on, or slightly above, a circular plane. We term this form of layout a pinboard layout, as the nodes placed slightly above the plane, and which are connected with the plane by edges, resemble pins stuck into a pinboard—see Figure 4 for an example.

This display represents two properties of the search term: relevance and popularity. Relevance is calculated as the degree to which the matching entity matches the search term. For example, given a search for the term “Macau” which yields an article with the title “Macau History”, this article would be assigned a score of 0.5, as half of its title matches the search term. Other nodes whose article text contains the search term would also be included in the search result. Nodes are positioned on the plane surrounding the search term node at a distance inversely proportional to their score. Moreover, nodes for which a match is found only in the article text and not in the title attribute are placed further away from the search term node than those nodes for which a match is found in the title attribute, as the presence of the search term in the title is considered to be a stronger determinant of relevance. The scores for presence of the search term in the title and text are cumulative, so that a node for whose article the search term is found in both title and text would have a higher score than a node for which the search term is found in only title or text, but not both (assuming an equal score for the titles and texts, respectively).

Popularity is recorded in the Wikipedia database as a counter of an entity’s viewership. Besides relevance, popularity is a useful property to consider when evaluating a search result. In the search result view, popularity is represented by elevating result nodes with a high popularity a small distance above the plane, connected with an edge to the point they occupy according to their relevance. The greater the popularity of an entity, the longer the edge is that elevates the corresponding node above the plane. This can be seen as the “pin” shaped nodes in the search result view of Figure 4.

The separation of the relevance and popularity properties that the pinboard layout visualizes enables the user to separately perceive these properties and to select search results based on the property they consider more important, rather than being presented with a result based on a combined relevance/popularity score, such as is commonly the case with modern web search engines like Google and others.

V. Conclusions

This paper has presented the WikiVis information visualization system for the Wikipedia system. It provides three different kinds of visualizations representing different relationships between entities in Wikipedia: entity views (using a star layout, based on the radial layout), category views (using nested cone tree layout, a modified form of the well-known cone tree layout [5]), and search result views (using a novel pinboard layout). These visualizations provide an understanding of relationships that are not easily perceived in the current Wikipedia system. This is of particular benefit to advanced Wikipedia users, including, but not limited to, researchers, librarians, and Wikipedia administrators, who are
seeking to better understand the complex relationships among articles and other entities.

Wikis in general, and Wikipedia in particular, are a recent phenomenon and to date there has been very little research into visualizing them. The work on history flow visualizations by Viegas et al. [6] has been specifically concerned with Wikipedia. However, it differs from the work presented in this paper by focusing on visualizing properties of a single article, while our visualization focuses on the relationships between articles as well as other Wikipedia entities. Work into visualizing the world wide web dates back to its early days (see [7] for a review from that time), but as it is concerned with the generic world wide web, not with the specific web-based application of Wikipedia, it differs from our work in two ways: its data source are web pages, whereas for WikiVis the data source is the Wikipedia database; and it determines web page relationships by examining the explicit link structure among pages, whereas in WikiVis the relationships are determined by examining co-author relationships between Wikipedia articles.

The work presented here is applicable beyond the Wikipedia system to which it was applied: Firstly, the visualization of article relationships is useful for wiki systems in general, as well as for web logs and other collaborative authoring systems. Secondly, the visualization layout methods presented are useful for visualizing other information structures with similar characteristics: large numbers of objects with differing degrees of relationships between them. For example, the layout methods could be applied to visualizing large document repositories such as those used in large corporate paperless office environments. The pinboard layout method could be usefully applied to supplement the current text-only search result views of web and other search engines by providing a secondary graphical search result view. Its advantage over the text-only view is the separate perception of relevance and popularity it enables in the search result, which allows users to select search results based on their priorities related to these properties.

This paper’s main contribution is the notion of using co-authorship for determining document similarity, which translates into an efficient and highly effective implementation. We expect this notion to be successfully applied to other application domains.

ACKNOWLEDGMENTS

Thanks are due to Stella Leong, Kathy Lou and Kitson Tong for their work on implementing the WikiVis system.

REFERENCES