A Method for Measuring Co-authorship Relationships in MediaWiki

Libby Veng-Sam Tang
Department of Computer and Information Science
Faculty of Science and Technology
University of Macau
Macau S.A.R., China
Tel: +853-8397 8635
libbyt@umac.mo

Robert P. Biuk-Aghai
Department of Computer and Information Science
Faculty of Science and Technology
University of Macau
Macau S.A.R., China
Tel: +853-8397 4365
robertb@umac.mo

Simon Fong
Department of Computer and Information Science
Faculty of Science and Technology
University of Macau
Macau S.A.R., China
Tel: +853-8397 4473
ccfong@umac.mo

ABSTRACT
Collaborative writing through wikis has become increasingly popular in recent years. When users contribute to a wiki article they implicitly establish a co-authorship relationship. Discovering these relationships can be of value, for example in finding experts on a given topic. However, it is not trivial to determine the main co-authors for a given author among the potentially thousands who have contributed to a given author's edit history. We have developed a method and algorithm for calculating a *co-authorship degree* for a given pair of authors. We have implemented this method as an extension for the MediaWiki system and demonstrate its performance which is satisfactory in the majority of cases. This paper also presents a method of determining an expertise group for a chosen topic.

Categories and Subject Descriptors
I.7.5 [**Document and Text Processing**]: Document Capture – *document analysis*, H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods*.

General Terms
Algorithms, Measurement, Experimentation.

Keywords
wiki, co-authorship, analysis.

1. INTRODUCTION
Working collaboratively in scholarly writing has been increasing in the past few decades [5]. The rapid development of computer-mediated communication systems facilitates and accelerates this working style around the world. Among the systems used for supporting collaborative writing, wikis have gained popularity and widespread use within the past few years. Wiki systems enable people working in different locations to communicate and share their expertise easily. Authors with an interest and expertise in a specific topic are enabled to contribute to writing on that topic. When writing on public wiki sites such as Wikipedia, however, co-authorship emerges as an implicit relationship from working on the same article, rather than being planned from the outset as is the case in traditional collaborative writing. Co-authorship of a paper can be thought of as documenting a collaboration between two or more authors [14].

Consider the situation depicted in Figure 1. The four circles represent authors *a*, *b*, *c* and *d*, and the two boxes represent articles *x* and *y* authored by them. The solid lines connecting authors to articles indicate that the given author has contributed to the given article. As shown, authors *a* and *c* have contributed to both articles *x* and *y*, whereas author *b* has only contributed to article *x* and author *d* has only contributed to article *y*. Thus there exists a co-authoring relationship among the authors of a given article, such as among the group of authors *a*, *b* and *c* in the case of article *x*. However, as we show later, this co-authoring relationship is not equally strong among all members of such a group. Therefore we analyze co-authorship relationships not on the group level but on the level of pairs of co-authors. In the example given there are five pairs of co-authors, i.e. a mutual co-authorship relationship exists for them, namely for the pairs (*a*, *b*), (*a*, *c*), (*a*, *d*), (*b*, *c*) and (*c*, *d*), represented by the dashed lines connecting authors. Out of these, the pairs (*a*, *b*) and (*b*, *c*) are for co-authoring of article *x* only, and the pairs (*a*, *d*) and (*c*, *d*) are for co-authoring of article *y* only. However, the pair (*a*, *c*) is for co-authoring both articles *x* and *y*. If we wish to determine how strong the co-authoring ties are for the given co-author pairs, the pair (*a*, *c*) should score higher than the other pairs based on the fact that its authors are jointly involved in twice as many articles.

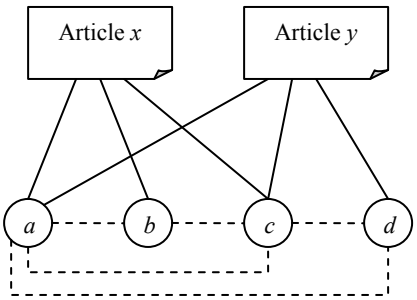

Figure 1. Co-authorship on two articles *x* and *y*

WikiSym '08, September 8-10, Porto, Portugal.

A naïve calculation would thus assign twice as high a value for co-authorship for this pair of authors. However, an author may have contributed to a given article just once or maybe many times. If, say, authors *a* and *b* had contributed to article *x* numerous times, but author *a* had contributed to article *y* only once, and author *c* had contributed to both articles *x* and *y* only once each, the strength of the co-authorship relation should be much greater for the pair of authors (*a*, *b*) than for (*a*, *c*). Other factors too should be taken into consideration, such as whether the co-authoring was concurrent or separated in time. Moreover, our discussion above was simplistic in assuming that the strength of a mutual co-authorship relationship for a given pair of authors *i* and *j* is identical in each direction, i.e. from *i* to *j* and from *j* to *i*. As explained later, this assumption is not true. In order to calculate the degree of co-authorship that takes all of the above factors into account we propose a new method in this paper.

Discovering co-authorship information can be useful in different areas. Significant co-authorship also implies co-expertise, and thus being able to discover the implicit relationship among authors in the wiki helps uncover expertise groups which can be of value when seeking experts in a given area. Based on our method for calculating the degree of co-authorship we have created such an application that determines a set of experts in a chosen area by searching for significant co-authors on articles within that area. Another application area is in automatic document classification which has been the subject of recent research in library information and science [16]. Measurement of degree of co-authorship has the potential to be applied in this area to help categorize documents, such as uncategorized articles in a wiki system. A further application area is for recommender systems where articles may be recommended to readers based on the co-authorship of an article being currently read.

The remainder of this paper is organized as follows: Section 2 reviews the related literature. Section 3 briefly reviews core concepts related to co-authorship. Section 4 presents an analysis of wiki data used by us. Section 5 then introduces the method and algorithm used in our method of calculating the degree of co-authorship. In Section 6 we present a case study of applying our method to the MediaWiki system, and in Section 7 we evaluate our method in an expertise finder application. Finally, Section 8 concludes this paper.

2. RELATED WORK

In recent years, the research problem of finding co-authorship has received much attention. Most of the work in the literature focuses on analyzing co-authorship by means of graphical visualization and using network analysis. Biuk-Aghai [1] presented a collection of display layouts for visualizing co-authorship networks in online Wikipedia. The visualizations display relationships between entities, between categories, and between search results, respectively. There are generally three schools of underlying techniques that support visualizing co-authorship, such as co-authorship network [3], [13], [14], [21], social network [18], and small-world network [19]. These techniques are based on the principles of statistics, graph theory, psychology and even a combination of them.

In particular, Huang [7] used social network algorithms to compute the co-authorship information into a custom-built InterRing visualizer for enabling users to understand the academic collaboration and knowledge domain of individuals from a computer science bibliography database. The purpose is to create a Research Quality Framework for assessing the research quality of individuals and research groups. Jdidia et al [8] enhanced the algorithm by using a social evolving graph that iteratively prunes less important arcs, together with a hierarchical clustering algorithm, on the articles from IEEE Infocom conference proceedings. They show that direct co-authors in the Infocom co-authorship network have a significant impact on the Program Committee board. Nascimento, Sander and Pound [12] tackled a similar problem on articles from the ACM SIGMOD community, using social networks instead. Another similar approach was adopted by Liu et al [10] where a weighted directional network model is used in which frequent collaborations are given higher weight, for enabling users to analyze graphs of co-authorship via a visualization toolkit. Goldenberg and Moore [4] studied co-authorship analysis on medical publications using Bayes-Net, claiming that it is relatively scalable, robust to noise, and supports query results with probabilities.

Nevertheless, the data-source used in the experiments of the above-mentioned articles mainly comes from DBLP, one of the largest academic databases of articles. The massive amount of data is downloaded to local storage for off-line analysis. Online performance by the graphical visualization techniques and network analysis was thus not of a concern. For instance, InterRing [7] is a visualizer that draws the results as widgets for visual inspection in a web version. It may not be suitably integrated into an online interactive system such as a wiki. Likewise in [10] the authors implemented the first version as a web application built by Webdot visualization tool, the second version as a standalone Java swing application based on TouchGraph. The other graph model implementations are offline/standalone analysis applications, as a specialized tool for the users (or analysts) who specifically want to study the co-authorship for visually gaining some insights from the models.

However, when working in an online setting such as in a wiki system, the algorithm for extracting the related authors by co-authorships must be relatively fast. Calculating the degree of co-authorship is an add-on feature for a wiki where users expect to retrieve a list of co-authors ranked by their intensity of collaboration with a particular author of a particular article. While in such an application scenario there is no requirement for extensive graphical displays of relations among co-authors, ease of use and fast speed are of greater importance.

In this situation, it is doubtful that graphics-intensive programs as documented in most of the literature could meet online response time requirements of general web users, whose tolerance threshold is around two seconds [11]. In this paper we present an alternative for the calculation of degree of co-authorship.

3. ARTICLES IN WIKI

We briefly review some key concepts related to wikis. This is based on the popular wiki application MediaWiki which we have chosen to apply our method to. However, the underlying concepts of our method are common in all wiki systems that maintain a revision history.

The primary entities in a wiki system are: *articles*, *revisions* and *authors*. The relationships between these entities are shown as an ER diagram in Figure 2. For each article there must be at least one

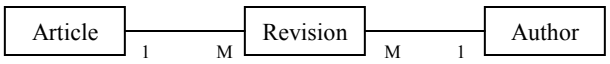

Figure 2. Relationship between article, revision and author

revision, but it is possible for an article to have many revisions. Each revision has exactly one author. Each author may have many revisions. Thus, there are *one-to-many* relationships between article and revision, and between author and revision.

3.1 Article

Wiki systems may contain many articles. Each article has a title and may also have other attributes. Each time an article is modified, a new revision is created. The latest revision constitutes the current state of the article. Different authors may modify the article, thus becoming co-authors. Wiki systems additionally usually provide the concept of categories for classifying articles, such as in the MediaWiki system.

3.2 Revision

Wiki systems maintain a history of an article's revisions which may be viewed and compared. In most wiki systems revisions are immutable and the entire revision history is permanently maintained. Revisions have attributes such as author, revision date and time, and other details. Revisions of an article may be retrieved from its history and the details of a revision viewed. When a new revision of an article is submitted, the submit time, user ID, and other revision information are stored. This revision history data is the main data used in our method.

MediaWiki further distinguishes two types of revisions: minor edit and non-minor edit. An author modifying an article in a MediaWiki system can indicate whether or not the edit is a minor one. Correcting spelling, grammar or punctuation are examples of minor edits, whereas adding paragraphs of new text is an example of a non-minor edit.

3.3 Author

Authors in a wiki are users who create or modify articles. A wiki system may allow anonymous users to author articles, or may restrict this to registered users only. Registered users have a unique user ID so the user can be identified as the same author. Anonymous users can typically only be identified by the IP address of their computer, and given that the same IP address may be assigned to different computers at different times, and that a given computer may be used by different users at different times, this identification effectively is useless in definitely identifying authors. For this reason, this paper mainly focuses on contributions from registered users, but also considers the influence of anonymous users.

If two registered authors in the wiki system are authors of the same article, that is, they have at least one revision each belonging to the same article, then a co-authorship relationship exists between them and they are considered co-authors. This kind of relationship is measured in our method.

4. DATA ANALYSIS

In developing our calculation method we used the popular wiki application MediaWiki and data from the well-known Wikipedia site. Wikipedia is a popular online encyclopedia which is written collaboratively by volunteers all around the world in more than 250 languages. General properties of the Wikipedia system and its user community have been analyzed in [15]. Other research has been done on analyzing article quality in Wikipedia and found that the cooperative content production model of Wikipedia results in high article quality [20]. The reasons why people contribute effort to Wikipedia without any financial return have been studied in [17], which revealed a multi-faceted picture of motivations, including mainly feelings of satisfaction in doing something useful, and enjoying sharing knowledge with others. Contributions to Wikipedia can also be categorized by type, such as adding new content, correcting grammatical and spelling mistakes, adding citations/references, and categorizing articles, etc. This diversity of contributions has resulted in the high quality of much of Wikipedia content.

Data dumps of Wikipedia databases are available for download. This enabled us to evaluate our method using actual data. The largest language editions of Wikipedia, English, German and French, now have more than 500,000 articles each. Their data sets are very large, measuring in the dozens of gigabytes, making the data difficult to handle and requiring large computing resources. A too small language edition, however, would not contain enough data to make a meaningful evaluation possible. We selected the Wikipedia Simple English language edition, which has a medium-size data set. As of 27 April 2008 it contained about 27,000 useful articles.

Article, revision and author are the three main conceptual elements involved in our method. In MediaWiki, article is called page and each page has a unique page ID[1]. Also, each revision has a unique revision ID. The revisions belonging to the same page each contain the same page ID. Each registered user has a unique user name and is assigned a unique user ID. Data related to revisions involved in our method is shown in Table 1.

Table 1. Revision

Column	Description
REV_ID	A unique revision ID of a revision.
REV_PAGE	The page ID of the revision belongs to.
REV_USER	User ID of the modifier.
REV_USER_TEXT	User name of the modifier.
REV_MINOR_EDIT	1 indicates this revision is only a minor edit and 0 indicates not.
REV_TIMESTAMP	The modify time of the revision.

The data dump used by us is the one from 16 November 2007. It contains 550,812 revision records in the revision table, belonging to 57,441 pages. The pages are categorized in different types by namespace. *Useful* articles are those which do not include user

[1] More specifically, an article is a page in one of the content namespaces, such as main, that is not a redirect and contains at least one internal link. See: http://www.mediawiki.org/wiki/Manual:Article . In the remainder of this paper we use the two terms interchangeably when the context is the MediaWiki system.

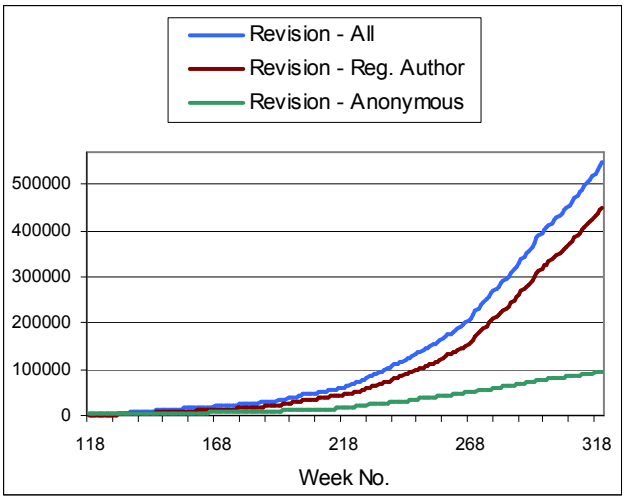

Figure 3. Number of revisions by week (all, registered, and anonymous authors)

pages, talk pages, help pages and page redirects. There are about 20,000 useful articles in this data dump. However, we use all pages, not only useful articles. About 82% of revisions are submitted by registered authors, the remaining 18% are from anonymous users. In total there are 5,401 registered authors. Of the revisions submitted by registered authors about 34% are non-minor edits. We calculated co-authorship degrees on a weekly basis from the beginning of the wiki data until the date of the data dump. There was no registered author before the 118th week. The data for the last week, the 322nd week, was not complete and only had data of three days. Thus, the analysis only used the data from the 118th week to the 321st week, a total period of almost 4 years.

Revision is the most important element in our method. The number of revisions over the analysis period is shown in Figure 3. As is shown, revisions by registered authors far outnumber those by anonymous users. At the last day of the data dump, there were only about 17.56% of revisions authored by anonymous users, thus about 4.5 times as many revisions were authored by registered authors than by anonymous users.

Table 2. Average number of revisions per page

Category of revision	Rev./Article	%
All	9.8362	100.0%
Non-minor edit	5.3583	54.5%
Minor edit	4.4782	45.5%
Submitted by registered author	8.1414	82.8%
Submitted by anonymous author	1.6951	17.2%

Over the analysis period, each article had an average of nearly 10 revisions. Table 2 shows the result of analyzing all data from the beginning to the last date in the data dump. Slightly over half of all revisions are non-minor edits. The difference between non-minor and minor edits is not significant, only about 9%. Over 4/5 of all revisions are submitted by registered authors. Moreover, the growth of numbers of revisions submitted by registered authors is

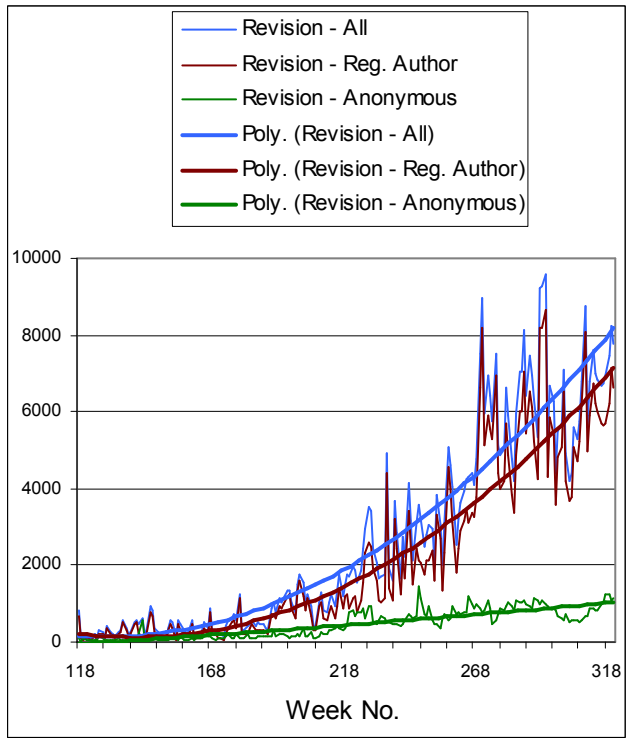

Figure 4. Increment of revisions by week (all, registered and anonymous authors)

faster than those from anonymous users, as shown in Figure 4. The curve of registered authors is close and similar to that of all revisions and we can infer that the growth of revisions by registered authors will follow the growth of all revisions, which in this case we observe follows an exponential growth path.

5. METHOD AND ALGORITHM

We have developed an algorithm for calculating the degree of co-authorship of a pair of authors. The input parameters are the IDs of two registered authors and the output is the degree of co-authorship. Revision information of the entire wiki system is used in this algorithm. All necessary information mentioned in this section can be retrieved from the revision table in MediaWiki. When an author submits a new revision, the required information is automatically recorded.

5.1 Method

Our calculation of degree of co-authorship initially filters author and article (page) data to eliminate authors and articles that are irrelevant for our calculation. For a given author, we consider other authors irrelevant as co-authors if they have only made minor edits on articles to which the given author has contributed; and articles as irrelevant if the given author has only made minor edits to these. This is to ensure that only significant co-authorship relationships are considered. For a given author *a*, the selection of co-authors and articles, and the subsequent calculation of co-authorship degree, is thus made as follows:

1. Obtain the set of all pages edited by author *a*

2. Eliminate the pages from the set of all pages for which author *a* has only made minor edits
3. For each remaining page, obtain the set of other authors
4. For each set of other authors, eliminate those authors who have only made minor edits
5. For each page's set of other authors, calculate a *page degree*
6. Calculate the *co-authorship degree* from all page degrees

The calculation of steps 5 and 6 is explained in detail below.

5.2 Degree

The output sought by our method is the *degree of co-authorship*. This is the result of the function $d(a, b)$ that maps from a given pair of authors (a, b) to a real-numbered value indicating the strength of the co-authorship relationship for the given pair of authors. The larger the degree value, the stronger the relationship, and thus the higher the likelihood that if one author contributes to an article the other author will also contribute to that article. The range of d is the interval $(0, \infty)$, i.e. $\{d: 0 < d < \infty\}$.

The function d is not symmetric, i.e. $d(a, b)$ is not necessarily equal to $d(b, a)$. The reason for this is that the degree is affected by the time factor, as explained below.

5.3 Co-Authorship Degree

Our method calculates the degree of co-authorship $d(a, b)$ for the relationship from author a to co-author b. As b is a co-author of a, they have jointly authored at least one article. The total number of jointly authored articles is denoted as t, where $\{t: 1 \le t < \infty\}$. The calculation of co-authorship degree is defined as:

$$d(a,b) = s \times \sum_{i=1}^{t} p(a,b)_i \qquad (1)$$

Page degree p is the degree of co-authorship of authors a and b on page i only (explained below). Thus, the total degree of co-authorship is the sum of all page degrees for authors a and b. A scaling constant s is used to tune the result to a range suitable for representation. In the case of two authors a and b who have not co-authored any article $d(a, b) = 0$.

5.4 Page Degree

Following the definition of 5.3, page degree p of page i for authors a and b is defined as:

$$p(a,b)_i = \left(\frac{\min(n_{ia}, n_{ib})}{n_i} + k\frac{\min(m_{ia}, m_{ib})}{m_i}\right) \times \frac{L_{ia} \cap L_{ib}}{L_{ia}} \qquad (2)$$

where

- n_i is the number of all non-minor edits of page i
- n_{ia} and n_{ib} are the numbers of non-minor edits of page i by authors a and b, respectively
- m_i is the number of all minor edits of page i
- m_{ia} and m_{ib} are the numbers of minor edits of page i by authors a and b, respectively
- k is a minor edit constant which affects the weight of minor edits
- L_{ia} and L_{ib} are the editing periods of authors a and b on page i, respectively

A detailed explanation of each factor follows.

5.4.1 n, m: Number of all non-minor, minor edits

These are the total numbers of all non-minor and minor edits (i.e. revisions), respectively. These numbers include both revisions authored by registered and anonymous authors of a given article (i.e. page). Although co-authorship of anonymous authors is not measured, they are also authors of a given article and therefore should be considered in the calculation of the page degree.

5.4.2 $min(n_a, n_b)$, $min(m_a, m_b)$: Minimum of non-minor, minor edits by authors a and b

The numbers of edits made by authors a and b are also separated into non-minor and minor edits. For each type (non-minor, minor) the smaller number of edits of that type among authors a and b is obtained. This is a measure of the extent of co-authorship of two authors on a given article. The rationale is that we wish to get a measure of the extent to which the two authors have actually *co*-authored, rather than individually authored, the given article. Thus, if one author performed 100 edits on the article and the other one only made 1 edit, the extent of collaborative contribution to authoring that article would be considerably smaller than if each had made 50 edits. We thus use the minimum of both values in our calculation.

5.4.3 $min(n_a, n_b) / n$, $min(m_a, m_b) / m$: Proportion of author contribution in total

This is the proportion of the extent of co-authoring of non-minor and minor edits by the two authors a and b in the given article, over the total number of revisions of that type of that article. It indicates the strength of co-authorship for these two authors relative to the contributions of other authors of the same article. This is to determine the significance of the two authors' contributions to the given article. Thus if the two authors are the only contributors to an article and have each made 100 edits, this contribution is significantly greater than the case where they each made 100 edits but other authors contributed 1000s of edits. The range of this ratio is the interval (0, 1], i.e. the upper bound is 1 if all revisions were submitted by these two authors only.

5.4.4 k: Minor edit constant

Non-minor edits are revisions considered to be more significant contributions to quantity or quality of an article whereas minor edits usually only make minor corrections such as fixing spelling, punctuation and grammar. In determining co-authorship of an article we consider non-minor edits to be of greater importance than minor edits and therefore assign the minor edit a lower weight in the calculation of page degree. The minor edit constant k can be set to a value in the interval [0, 1] to give minor edit co-authorship a value ranging from no weight at all to equal weight with non-minor edit co-authorship. We use a value of $k = 0.05$, i.e. minor edits have only 5% of the weight of non-minor edits, a value that appeared to us to be reasonable considering the low significance of edits labeled as minor in Wikipedia.

5.4.5 *L: Authoring period of a given author*

For someone to be considered an author of an article, we consider not only the portion of their contributions in terms of revisions made by them, but also the length of time over which they have contributed to an article relative to the length of time this article has been in existence. That is, if an author makes contributions to an article only during a short period it will be considered less than if that author had made contributions over a longer period of time. The rationale is that a given author's influence on an article exists only during the time period to which they contribute to it. In the case of a singly authored work, the author is author of that work over its entire duration. However, in wikis where authoring is shared different authors may exert different influence over an article by participating in or withdrawing from editing it. Therefore we consider the authoring period as another important factor in calculating co-authorship degree for a given article. An author's first non-minor edit on a page is considered the beginning of the authoring period, and the last non-minor edit of this author is considered as the end of the authoring period. Thus the authoring period is the time interval between these two bounding points, measured in number of days.

5.4.6 *$L_a \cap L_b / L_a$: Proportion of intersection of authoring period over first author's authoring period*

In conventional co-authoring, such as on a book or journal article, the authors are usually engaged in the authoring process during more or less identical time periods. During this period the authors interact with each other to discuss jointly authored content, and with each other's writing by making revisions to it. In wikis, however, authors are not necessarily engaged in authoring during the same time periods, and it could be that one author starts contributing to an article after another author has finished his contributions to that article. In this case these authors thus do not interact with each other (such as through a wiki article's discussion page) and have no mutual interaction with each other's writing (although one user may change the other's writing, the converse will not be true). Therefore, we consider the intersection of the authoring periods of two authors when calculating their degree of co-authorship: the longer the intersecting period, the greater the degree of co-authorship as the authors would have been involved in joint authoring during at least part of that article's overall authoring period.

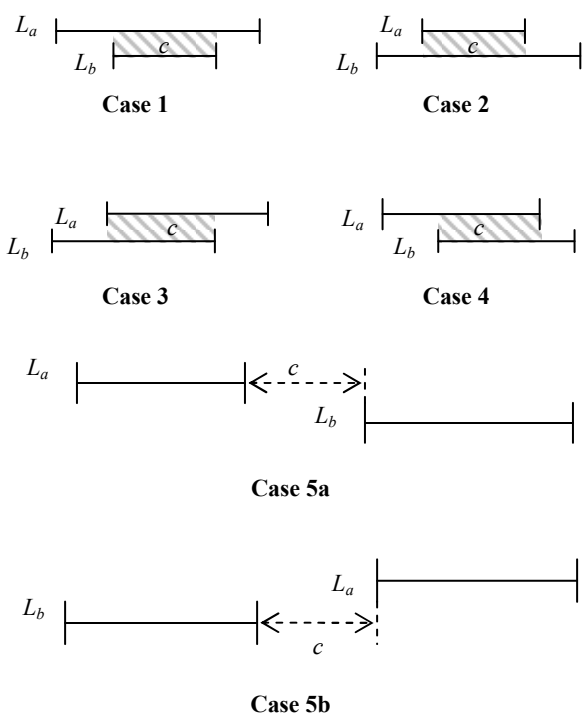

Figure 5. The five cases of $L_a \cap L_b$

The calculation of $L_a \cap L_b$ determines the length of the joint authoring period. There are five different cases of intersection, as shown in Figure 5: two cases where the intersection is equal to one of the two authoring periods (cases 1 and 2), two cases where the intersection is shorter than both of the two authoring periods (cases 3 and 4), and one case where the two authoring periods do not overlap, divided into two sub-cases (5a and 5b). Time is shown along the horizontal dimension, and the authoring periods of authors a and b is represented by the two parallel lines, that of L_a above that of L_b. In cases 1 through 4, the authoring periods L_a and L_b overlap. This overlapping period is the common authoring period and is denoted as c, shown as the shaded area. The value of c is an integer and is counted in days. For cases 1 through 4 this value is greater zero. In case 5, there is no intersection between two authors. However, non-overlapping periods of authoring should not be entirely disregarded, but should be considered with a reduced weight. Therefore we use the absolute value of c, which in case 5 is the distance between the end point of the earlier authoring period and the start point of the later authoring period, to calculate $L_a \cap L_b$ as $1 / |c|^{0.5}$. Thus a longer distance results in a smaller intersection value. The value of $L_a \cap L_b$ for $c > 0$ is c, and for $c < 0$ it is in the interval (0, 1].

Finally, the intersection value is divided by L_a. The resulting value indicates the proportion of a's authoring period that both a and b were working on the article. This part of the calculation results in a different page degree value for authors a and b, i.e. usually $p(a, b)_i \neq p(b, a)_i$ as the denominator value is different based on author. For instance if author a had been involved in an article for one year, and author b for only one day, then author b would only be a very insignificant co-author to a, but author a would be a very significant co-author to b. The value of $L_a \cap L_b / L_a$ is in the interval (0, 1].

5.5 Boundary

The upper bound for the page degree $p(a, b)_i$ is given when all contributions to a page are made by authors a and b only, in equal proportions both for minor and non-minor edits, and with identical authoring periods. That is, referring to equation (2), when $n_{ia} = n_{ib}$, $n_{ia} + n_{ib} = n$, $m_{ia} = m_{ib}$, $m_{ia} + m_{ib} = m$, $L_{ia} = L_{ib}$. Thus $p(a, b)_i$ will be:

$$p(a,b)_i = (\frac{1}{2} + k\frac{1}{2}) \times 1 = \frac{1+k}{2}$$

where k is the minor edit constant.

Therefore, referring to equation (1), the upper bound of co-authorship degree $d(a, b)$ is given when all page degrees are at the upper bound, that is:

$$d(a,b) = s \times \sum_{i=1}^{t} \frac{1+k}{2} = \frac{s \times t \times (1+k)}{2}$$

where s is the scaling factor and t is the number of pages co-authored. The values of s and k can be adjusted as desired. Thus, the upper bound of the co-authorship degree is determined by the value of t, i.e. how many pages were co-authored, which in turn is bounded by the number of pages in the system. Thus, as stated above, the range of d is the interval $(0, \infty)$.

5.6 Complexity

To evaluate the complexity of our algorithm, we firstly outline the overall steps of determining the degree of co-authorship for a given author:

1 FOR each co-author in the author's co-author list
  2 FOR each co-authored page authored by the given author and co-author
    3 CALCULATE the page degree for the current page of the given author and co-author
    4 ACCUMULATE the page degree as co-author degree
  END FOR
  5 STORE the co-author degree of the co-author
END FOR

There are five operations, among which operations 4 and 5 can be neglected since they are simple operations. The processing time of operation 3 approximates to constant. Refer to equation (2) above for calculating the page degree, k is a constant and the parameters m and n are the total number of revisions of the specified page under different conditions. Thus, the number of revisions of the specified page does not significantly affect the processing time for finding m and n in the process, and similarly for the time factor L. Therefore, the processing time for calculating the page degree is quite stable and only minimally influenced by the number of revisions. The remaining operations are the two loops which depend on the number of co-authors n and the number of co-authored pages mi of a given pair of authors. They are the main factors that affect the processing time significantly.

As the processing time of finding the page degree approximates to constant, the processing time needed for the algorithm can be simplified as follows:

$$\sum_{i=1}^{n} PT(d(a,b_i)) = \sum_{i=1}^{n}\sum_{j=1}^{mi} PT(p(a,b_i)_j) \approx \sum_{i=1}^{n}\sum_{j=1}^{mi} C$$

where

- $PT()$ is a function which returns the processing time
- a is the given author
- b_i is co-author i of the given author a
- $d()$ is the degree of co-authorship of a and b_i
- $p()_j$ is the page degree of page j co-authored by a and b_i
- C is a constant processing time needed for finding the page degree
- n is number of co-authors of a
- mi is number of co-authored pages by a and b_i, it is variant for each co-author pair i

Figure 6. Special page list with the Co-Authors extension

Assuming an average case of the number of co-authors n that is close to the number of co-authored pages mi, our algorithm thus has quadratic complexity, i.e. $O(n^2)$.

6. CASE STUDY

We have implemented our method and algorithm on the MediaWiki system and using data from Wikipedia.

6.1 Implementation of the Algorithm

We implemented our algorithm in PHP code as a MediaWiki extension, together with MySQL stored procedures. Extensions are self-contained pieces of PHP code that add new features or enhance functionality of the main MediaWiki core and that can be easily integrated in MediaWiki. Extensions are divided in different categories. The one we developed belongs to the 'Special page' category which is about adding new reporting and administrative capabilities. All special pages enabled in MediaWiki are listed by clicking the link 'Special pages' in the toolbox section of the left menu in MediaWiki. A list of special pages including our extension is shown in Figure 6. With our extension a MediaWiki user can discover the list of co-authors of a specified author by entering a username and desired sort order as shown in Figure 7. The co-authoring degrees are calculated for all co-authors according to our method as described in the previous section. The resulting co-author list displays username and co-authorship degree as shown in Figure 8. A bar whose length is proportional to the degree is included as a simple visualization to facilitate the comparison of co-authors. Moreover, for each listed co-author, a link "(co-authors)" is included after

Figure 7. Author search in the co-authors page

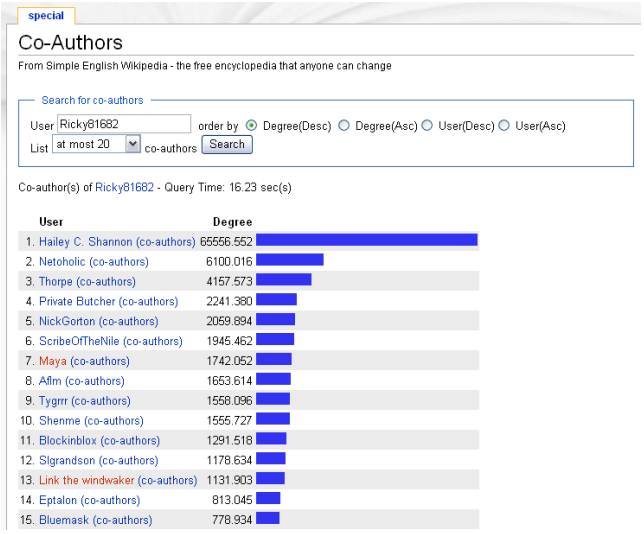

Figure 8. List of co-authors of a selected author

their username to allow a search for that co-author's co-authors.

Having implemented the algorithm we validated the results produced by our implementation by comparing them with manually calculated results. As the data volume involved tends to be very large, making it unwieldy for manual calculation, we chose several authors with a small number of pages, revisions and co-authors. The manually calculated results were identical to those produced by our implementation, leading us to believe that the results are valid also in the case of larger data volumes.

6.2 Performance

Our implementation was deployed on a Windows PC with a 3 GHz Intel Pentium 4 CPU, and 2 GB RAM. The Simple English Wikipedia database we used had 550,812 revision records. Through several rounds of trials and optimization we improved the performance so that it now is within an acceptable range for most user records. A previous study found that a waiting time of 2 seconds was tolerable for web applications [11]. Taking this as our primary performance target, we also set a secondary performance target of 10 seconds as the maximum acceptable waiting time. We then determined the distribution of the co-author calculation time relative to these two performance targets on our trial database, which is shown in Table 3.

Table 3. Distribution of calculation time *t*

$t \leq 2$ sec.	2 sec. $< t \leq 10$ sec.	$t > 10$ sec.
92%	7%	1%

Given the quadratic complexity of our algorithm, an implementation which calculates the degree of co-authorship in real time, such as ours, is only suited for wiki databases of moderate size. In the case of the Simple English Wikipedia database that we used there were about 5,400 registered authors and about 27,000 pages, and we achieved an acceptable real-time performance. However, in the case of much larger databases, such as the standard English Wikipedia one, which had about 2.5 million pages and 7.5 million registered authors as of July 2008, we expect performance of a real-time co-authorship calculation to be unacceptable. In such cases an offline pre-calculation performed periodically in batch mode would be more appropriate.

7. EVALUATION

The MediaWiki system and its extensions do not currently have a function to show co-authors of a given author. The case study introduced above demonstrates how to apply our method practically in the MediaWiki system. The applicability of our method is not limited to the MediaWiki system, however. Besides other wiki and co-authoring systems it can also be used in visualization systems such as WikiVis [1] which discovers implicit relationships among articles in Wikipedia. WikiVis is based on a simple notion of co-authorship between articles. By using our method, it will be possible to determine the strength of the relationship among Wikipedia articles more accurately. In other related work [2], the co-revision network of an article was used to find the similarity of pages and the concept of a co-author network was introduced. Again, our method will be able to contribute by allowing a more accurate calculation.

Another application of co-authorship is in finding expertise groups. When searching for experts on a certain topic, a search for co-authors of an article on the chosen topic can provide a good starting point. We outline below the process of finding expertise groups, as illustrated in Figure 9. The steps are:

1. Select a page or category closely related to the required area of expertise.
2. Find the main authors of the page or category.
3. Find the close co-authors of those main authors, i.e. with strong degree of co-authorship.
4. Apply rules to include selected co-authors in the group of main authors.
5. The resulting group of authors forms the expertise group of the selected page or category.

7.1 Main authors of page/category

Initially we define the main author of a page as the author who

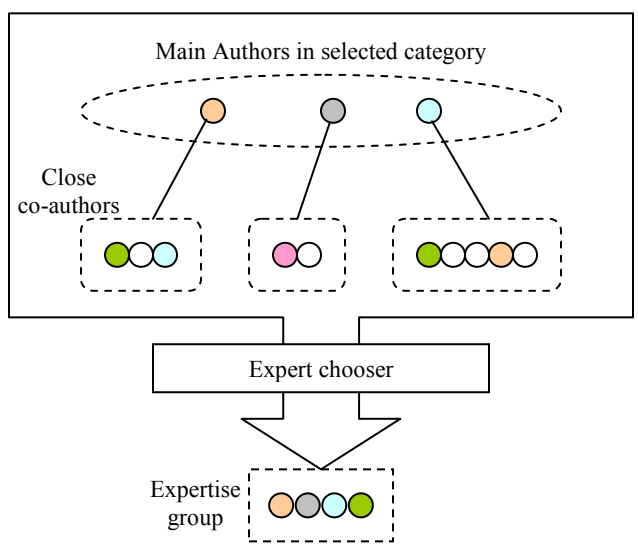

Figure 9. Finding the expertise group

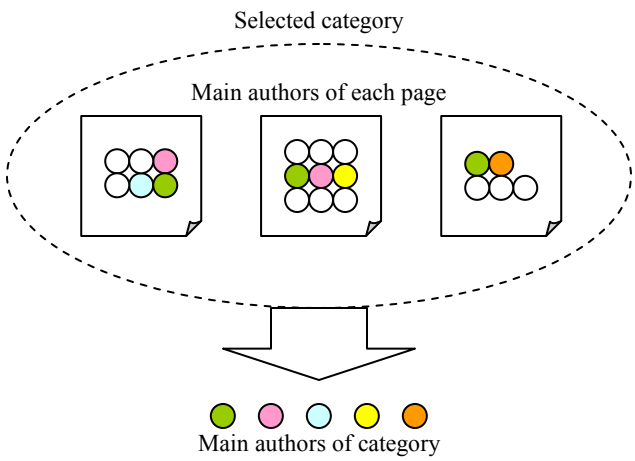

Figure 10. Finding the main authors of a category

has made a number of revisions that constitute a significant portion of all revisions. For instance, we may define that any author who has made 10% or more of all revisions is a main author of the given page. Alternatively we may define that the top *n* contributors of a page are the main authors of that page. Following the selection of main authors of a page, we can define a similar rule for the category. That is, a user who is among the main authors of a significant portion of all pages belonging to that category is defined to be a main author of that category. For instance, a user who is main author of 10% or more of all pages of a category is a main author of that category, or again an alternative selection could be to select the top *n* among the main authors of pages in that category, i.e. those users who are main authors of the largest number of pages of that category. The concept is represented in Figure 10.

7.2 Close co-authors of main authors

The close co-authors of main authors can be found using the method defined in this paper, namely by calculating the degree of co-authorship between them. Then, only the co-authors who fulfill a defined condition (such as degree of co-authorship > x, or top n co-authors) are considered as close co-authors, i.e. having strong ties of co-authorship with the main author.

7.3 Expert selection

Expert selection aims to determine the members of the expertise group. The main author group and the groups of close co-authors are analyzed and members chosen according to certain defined rules. For instance, a member of the main authors group who is also a member of at least n of the other main authors' groups of close co-authors is considered a member of the expertise group. Authors who are not main authors but are members of at least m main authors' groups of close co-author groups may likewise be considered members of the expertise group. This process is illustrated in Figure 9.

7.4 Example

We give a demonstration of the expertise group selection process, again for the Simple English Wikipedia data mentioned earlier and choosing the category 'Music' as an example. The first step is finding the main authors in this category. We implemented a

Figure 11. The implemented special page Main Authors

MediaWiki extension for finding the main authors of a specified category. The extension is similar to the co-author extension implemented before. A search function which accepts a category name, ordering and record limit of the result is provided as a special page, as shown in Figure 11. The result is a list of main authors with a page number indicating for how many pages in the given category the shown author is a main author. We used the criteria of registered users who have made at least 10% of all revisions for selecting the main authors. Likewise, for the category we defined that an author who is a main author of at least 10% of all pages of that category is also a main author of that category. System administrators and bots are not counted in this selection because their modifications are assumed as non-content related. In the example of the 'Music' category there are 126 pages with a total of 1,483 revisions. Out of these, 251 revisions were submitted by anonymous users and 705 revisions by administrators and bots, and are thus not included. The remaining 527 revisions were submitted by 124 distinct registered authors. Table 4 shows the main author list returned for this category.

Table 4. Main author list of category 'Music'

Author	Main author of number of pages	% of total
Hikitsurisan	56	44.44%
Zephyrad	13	10.32%

These two category main authors are initially selected as members of the expertise group. Subsequently we look for the members of the expertise group among the close co-authors of these main authors. We decided to use a degree of co-authorship of at least 0.1 to define a close co-author, and a criterion of the same author being a close co-author of at least two category main authors to select expertise group members. In this example, there are 28 co-authors of main author Hikitsurisan, and 10 co-authors of main author Zephyrad with degree ≥ 0.1, i.e. their close co-authors. Of these, three are co-authors of both of these two main authors. Therefore, the expertise group of the category 'Music' consists of its two main authors and these three close co-authors.

We have demonstrated how to determine an expertise group of a given article category in MediaWiki-based systems. Are members of this expertise group really experts in the selected category? This question begs further research. However, we are confident that our method provides a good starting point in this direction.

8. CONCLUSIONS

Wikis have gained increasing importance and use throughout the world. Through their use, implicit groups of expertise are established around different topics. However, as these relationships are difficult to discern, data analysis techniques are used for discovering them. In this paper we have presented a new method for calculating the degree of co-authorship for a given pair of authors. This method is more accurate than any other existing methods that analyze co-authorship, and has satisfactory performance that makes it suitable for online use. Besides use in the MediaWiki system, it has the potential to be integrated in other collaborative writing systems that maintain a complete edit history. For instance, a wiki in a large organization can reveal expertise groups that are not explicitly recorded or known. Our method also has the potential to be integrated in visualization systems that display relationships of wiki entities.

Future work will better characterize revisions. In the current algorithm we only distinguish between minor and non-minor edits (which are explicitly labeled as such by the author). Our next step is to analyze revisions made to be able to distinguish, for example, between new additions of text, spelling/grammar correction, editing to bring an article in line with style guides, and other types of edits. A different weight can then be assigned to each type and used in the calculation of co-authorship to produce a more accurate result, rather than using a fixed ratio as with the minor edit constant in our current algorithm.

9. ACKNOWLEDGMENTS

The financial support from the University of Macau Research Committee is gratefully acknowledged.

10. REFERENCES

[1] Biuk-Aghai, R. P. 2006. Visualizing Co-Authorship Networks in Online Wikipedia. Communications and Information Technologies, 2006. ISCIT '06 (Bangkok, Thailand, September 2006), 737-742.

[2] Brandes, U. and Lerner, J. 2007. Revision and co-revision in Wikipedia. Proceedings of the International Workshop on Bridging the Gap Between Semantic Web and Web 2.0 at the 4th European Semantic Web Conference (ESWC'07) (Innsbruck, Austria), June 7, 2007, 85–96.

[3] Chen, C., Paul, R. J., Visualizing a Knowledge Domain's Intellectual Structure, IEEE Computer. 34(3), March 2001, 65-71.

[4] Goldenberg, A., Moore, A. W., Bayes Net Graphs to Understand Coauthorship Networks?, LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery, August 2005, 1-8.

[5] Hart, L. 2000. Co-authorship in the academic library literature: A survey of attitudes and behaviors. The Journal of Academic Librarianship, 26(5):339-345, September 2000.

[6] Hart, L. 2007. Collaboration and Article Quality in the Literature of Academic Librarianship. The Journal of Academic Librarianship, 33(2):190-195, March 2007.

[7] Huang, T.-H., Huang, M. L., Analysis and Visualization of Co-authorship Networks for Understanding Academic Collaboration and Knowledge Domain of Individual Researchers, 2006 International Conference on Computer Graphics, Imaging and Visualisation, 26-28 July 2006,18-23.

[8] Jdidia, M. B., Robardet, C., Fleury, E., Communities detection and analysis of their dynamics in collaborative networks, 2nd International Conference on Digital Information Management, 2007. ICDIM '07, Volume 2, 28-31 Oct. 2007, pp.744 – 749.

[9] Ke, W., Borner, K., Viswanath, L., Major Information Visualization Authors, Papers and Topics in the ACM Library, IEEE Symposium on Information Visualization, 2004. INFOVIS 2004, 10-12 Oct. 2004, r1 - r1.

[10] Liu, X., Bollen, J., Nelson, M. L., Van de Sompel, H., Hussell, J., Luce, R., Marks, L., Toolkits for Visualizing Co-Authorship Graph, Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004, 7-11 June 2004, 404.

[11] Nah, F. 2004. A study on tolerable waiting time: how long are Web users willing to wait? Behaviour and Information Technology, Volume 23, Number 3, May-June 2004, 153-163(11).

[12] Nascimento, M. A., Sander, J., Pound, J. Analysis of SIGMOD's CoAuthorship Graph, ACM SIGMOD Record, Volume 32 Issue 3, September 2003, 8-10.

[13] Newman, M., Scientific Collaboration Networks: I. Network Construction and Fundamental Results, Physical Review E, 64(1):016131, 2001.

[14] Newman, M. E. J. 2004. Coauthorship networks and patterns of scientific collaboration. P NATL ACAD SCI USA, 101(1):5200-5205, April 2004.

[15] Ortega, F. and Barahona, M. 2007. Quantitative Analysis of the Wikipedia Community of Users. In Proceedings of the 2007 International Symposium on Wikis (Montreal, Quebec, Canada, October 21-23, 2007), 75-86.

[16] Pong, Y., Kwok, C., Lau, Y. Hao, J. and Wong, C. 2008. A comparative study of two automatic document classification methods in a library setting. Journal of Information Science, Volume 34, Issue 2 (April 2008), 213-230.

[17] Prasarnphanich, P. and Wagner, C. 2008. Creating Critical Mass in Collaboration Systems: Insights from Wikipedia. IEEE DEST 2008 (Phitsanulok, Thailand, Feb. 26-29, 2008), 126-130.

[18] Wasserman, S., Faust, K., Social Network Analysis, Cambridge University Press, Cambridge, 1994.

[19] Watts, D. J., Strogatz, S. H., Collective dynamics of 'small-world' networks, Nature, 393:440–442, 1998.

[20] Wilkinson, M. and Huberman, A. 2007. Cooperation and Quality in Wikipedia. In Proceedings of the 2007 International Symposium on Wikis (Montreal, Quebec, Canada, October 21-23, 2007), 157-164.

[21] Yoshikane, F., Nozawa, T., Tsuji, K., Comparative Analysis of Co-authorship Networks Considering Authors' Roles in Collaboration: Differences between the Theoretical and Application Areas, ISSI 2005, July, 2005, vol.2, 509-516.