Clustering of Wikipedia Pages on Edit Behaviors

Nilothpal Talukder, Haiqiong Li, Malik Magdon-Ismail
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY, USA
{talukn,lih9}@rpi.edu, magdon@cs.rpi.edu

***Abstract*—We consider the edit history of Wikipedia to perform clustering of the pages. We conjecture that the editors exhibit homophily or high correlation (in terms of the topics of interests). Therefore, it is possible to utilize the edit history to cluster pages having same or closely related topics. We validate our clustering results with the list of categories and the incoming and outgoing links on the Wikipedia pages. We use *k*-means to perform page clustering. Typically, Wikipedia page editors demonstrate multiple interests and Wikipedia pages list multiple categories, whereas *k*-means delivers only partitioning. Therefore, we also study the results from a clustering algorithm called "Connected Iterative Scan" that produces overlapping communities from a social graph. We also study page dynamics which can potentially be incorporated into the clustering of pages.**

Keywords-Wikipedia page edits, collaboration graph, clustering

I. INTRODUCTION

It goes beyond saying that Wikipedia provides a wealth of information in today's internet age. This is a great collaborative effort among the volunteers all around the world that resulted in such a rich source of accumulated knowledge. The inception of the effort dates back to 2001 and since then it is gaining tremendous attention every year. Besides the main content, each wiki page keeps track of the timestamps of edits and identities of the editors. The identity is stored in the form of user ID and/or IP address. By performing edits on the pages, editors are actually participating in a social process. The editors having similar interests are likely to perform edits on the pages that have the same or very closely related topics.

We study the social process by a weighted graph (collaboration graph) which encodes the "edit behavior" or the collaboration among the editors. In the collaboration graph, each editor is a node, and edges between editors reflect their edit history on the same pages. We conjecture that the "edit behavior" can be utilized to find "good" clustering of the wiki pages having same or very closely related topics. This is because the editors exhibit high correlation or homophily [1] on the collaboration graph due to similar interests. The notion of similar interests can be easily reflected into the similar topics while clustering the wiki pages. Finding similar Wikipedia pages has many potential applications. For example, wiki pages that possibly belong to the same topic can be recommended to the reader to enhance user experience.

Wikipedia maintains the category of pages that are structured but ill-defined. Each Wikipedia page actually belongs to multiple categories, thus the organization of pages in Wikipedia does not form a tree structure. Instead, it looks more like an acyclic directed graph [2]. As of September 2011, Wikipedia has around 1.2 million categories, whereas the total number of pages is about 4 million [3]. Nevertheless, we use this category information to validate our primary findings on the clustering of the pages using Jaccard Index. In addition to categories, page link information among Wikipedia pages is also used for computing the similarities between pages. These similarities are used to further validate our result. We use the terms "page" and "article" interchangeably throughout the paper.

The contributions of our paper are the following:

1) We study both overlapping and non-overlapping clustering of the Wikipedia pages based on the homophily of edit behavior.

2) We validate our clustering results with both original Wikipedia category data and page-link based similarity.

The remainder of the paper is organized as follows. Section II discusses the preliminaries. Section III presents our approach on clustering the Wikipedia pages. Results and validation are discussed in Section IV and Section V respectively. Finally, Sections VI, VII, and VIII cover related work, additional discussions, and future work.

II. PRELIMINARIES

In this section we introduce the definitions and the notations used in the paper.

A. Colloboration Graph

The weighted collaboration graph, encodes the interaction among the editors. A weighted edge $G_c = (V_C, E_C)$ with the weight w_{ij} denotes that two distinct editors $v_i, v_j \epsilon V_C$ have performed w_{ij} edits together (on the same page) during a given timeframe. The threshold value θ denotes that the editors edited the same page at least θ times. A collaboration graph, $G_c = (V_C, E_C, \theta)$ with the threshold θ indicates that $\forall w_{ij} \epsilon\, E_C, w_{ij} \geq \theta$.

B. Connected Iterative Scan (CIS)

Standard clustering algorithm delivers disjoint sets of vertices as clusters. In other words, there is no overlap among the vertex sets. However, in social network the communities often have large overlaps. It is obvious to see that the

communities of editors with similar "interests" can have such overlaps due to interests in multiple topics by an editor. In order to find overlapping communities in a social graph we can have three major criteria that govern the process.

1) *The internal edge density should be high.*

2) *More communication internally than outside.*

3) *A community should be "locally" optimal*

The second criterion necessitates that the number of edges going outside of the community (E_{out}) should be smaller than the number of edges inside the community (E_{in}). It turns out that this criterion is a weak one. A community can have lots of interactions outside of it for which $E_{out} \leq E_{in}$ does not hold. On the other hand, finding all the subsets, *S* that satisfies "weak community criterion" is a hard problem.

The Connected Iterative Scan (CIS) [7] algorithm provides a heuristic for extracting overlapping communities from the social graph. The criteria are modified into three sets of axioms as the following:

1) Any definition should be "local"

2) Any community should have enough E_{in} so that it is connected

3) The community quality or the measure of density, $d(S)$ is locally optimal, where:

$$d(S) = \frac{E_{in}}{E_{in} + E_{out}}$$

During each iteration of the algorithm, all vertices are examined and the vertex that causes the maximum increase of the density is picked. The current set of nodes representing the community is updated with that vertex. It can be shown that a long chain of small degree nodes will always increase the density and result in a long chain of sparse community. A parameter λ was introduced to restrict the extent of the communities. The density function with λ is defined as following:

$$d(S) = \lambda \frac{E_{in}}{E_{in} + E_{out}} + (1 - \lambda) \frac{2 \times E_{in}}{|S| \times (|S| - 1)}$$

Setting $\lambda = 1$ will ignore the restriction on the community size. We chose $\lambda = 0.4$ to identify communities from the collaboration graph.

III. APPROACH

In this section, we discuss the methods we use to cluster Wikipedia pages into similar clusters. We first pre-process the wiki page data to reduce data size. Then, we perform clustering on the filtered data to obtain wiki page clusters.

A. Pre-processing

Since Wikipedia is a free, web-based, collaborative encyclopedia project, almost all of its 20 million articles can be edited by anyone with access to the site. This results in a very large number of pages and editors. The Wikipedia editing dataset we used consists of the edits performed by a list of editors on weekly granularity. To allow efficiently performing the clustering of the pages based on the edit data, we need to reduce the data to a reasonable size. We generate collaboration graphs from the edit history between the years 2001 to 2008, one graph for every year. The threshold value θ was chosen to be 5. That means the editors considered in the graphs have performed at least 5 edits together on a page in the same year. Since we validate our findings with the Wikipedia category data, we consider the pages with no categories as spurious pages and filtered them out from the collaboration graphs. These spurious pages are mostly user talk pages instead of Wikipedia articles.

We further process the data and only keep the *P* most frequently edited pages and the *U* most active users within each graph. For each Wikipedia page *p*, we scan the editing data and find the number of editors who edited the *p* in a particular year. Then we obtain the top *P* pages that have the most editors. We rank all the editors based on the number of distinct articles they authored in a particular graph and then find the top *U* editors.

B. Clustering of Pages using Editor Data

We use the standard *k*-means clustering algorithm to cluster Wikipedia pages. The standard *k*-means states that given a set of *n* observations ($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\cdots$, $\mathbf{x}_n$), each of which is a *d*-dimensional vector, the algorithm partitions the observations into *k* clusters C = {C_1, C_2, C_3, …, C_k} so as to minimize the intra-cluster distance:

$$\underset{\mathbf{C}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\mathbf{x_j} \in \mathbf{C_i}} \left\| \mathrm{x_j} - \mathrm{centroid_i} \right\|^2$$

We use the authorship information of each page to define the distance between pages. Our hypothesis is that two pages that have similar authors should be similar to each other and possibly belong to the same category. This is because these authors share similar interest and are likely to collaborate on a series of pages covering the same topic.

To cluster *P* pages, we create a observation vector for each page: ($\mathbf{x}_1$, $\mathbf{x}_2$,…, $\mathbf{x}_P$). Each of these vectors is *U*-dimensional. Each element in the vector corresponds to one of the *U* most active editors. If a page $\boldsymbol{p}_i$ has been edited by a certain editor *j* ($j \in [1 \ldots U]$), then the *j*-th element in vector $\mathbf{x}_i$ is set to 1. The distance between Wikipedia pages are defined as distances between these binary vectors.

An alternative scheme is to use the number of edits by each author instead of binary decision in the observation vector. For example, if page $\boldsymbol{p}_i$ has been edited by editor *j* for *m* times, then the *j*-th element in vector $\mathbf{x}_i$ is set to *m*. In experiments, we found this is less effective than using binary values because higher weights are given to those editors who tend to edit a page in multiple editing sessions. This additional layer of editing behavior weakens the information we hope to capture—the interests of each editors rather than the editing frequency.

After creating the *P* observations of the most frequently edited pages, we perform *k*-means clustering with a set of

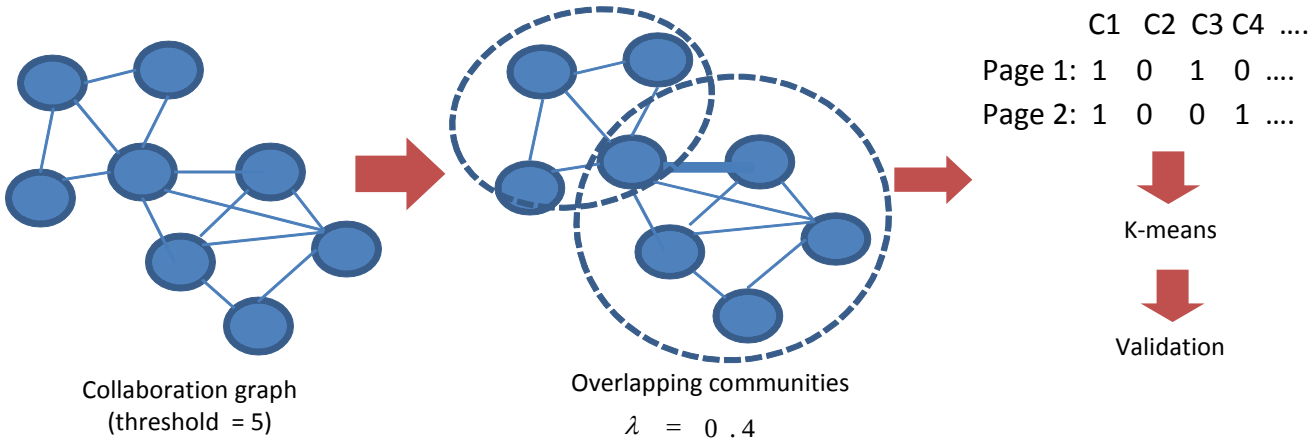

Figure 1 Steps involved in the community based editing behavior extraction and validation.

selected number of k values. We also performed outlier rejection to reject clusters that are too small (e.g., 1 -2 page) and clusters that their intra-cluster distance is larger than a certain threshold.

C. Clustering of Pages using Editor Community

Another way to cluster Wikipedia pages is to use editing behavior in terms of editor community instead of individual editors. For this purpose, we first run the Connective Iterative Scan (CIS) algorithm, as outlined in Section II, on the collaboration graph to partition editors into C communities. Each editor can belong to more than one community.

After computing the community for each editor, we compute the observation vectors for each page. These vectors are C-dimensional. Each element in the vector corresponds to a community ID. If a page $\boldsymbol{p}_i$ has been edited by any editor belonging to community j ($j \in [1 \ldots C]$), then the j-th element in vector $\mathbf{x}_i$ is set to be 1. We perform the k-means clustering the same way as in the previous sub-section. This process is shown in Figure 1.

IV. Validation

Once the clusters of the Wikipedia pages are identified, we attempt to find the "meaningfulness" of the clusters. In other words, we examine how similar or closely related the pages inside a cluster are. Then we compare our results with the results from two different random graph models on category and page link data.

To determine the similarity between two Wikipedia pages, we consider the fraction of common categories and page links on the pages. Cluster quality is measured in terms of the average intra-cluster distance of the clusters based on these two criteria. In this section we describe our cluster quality measures and random cluster generation approaches.

A. Wikipedia Category System

As of today, Wikipedia consists of 1.2 million categories. Each page belongs to many categories. For example, the current Wikipedia article for "Albert Einstein" consists of some of the following categories out of total 64 entries: *Albert Einstein, 1879 births, 1955 deaths, 19th-century German people, Academics of Charles University in Prague, American philosophers, American theoretical physicists, Cosmologists, Deaths from abdominal aortic aneurysm, Einstein family, ETH Zurich alumni, ETH Zurich faculty*.

Figure 2 highlights the common categories on the "Facebook" and "Myspace" Wikipedia articles. We observe that there are six categories in common. The categories are mostly created and listed on the articles by the editors at different times. It can be observed that an article may not list all the relevant categories it should ideally belong to. For example, the categories "internet advertising and promotions" and "virtual communities" are listed on the "Myspace" page. However, they certainly apply to "Facebook" page, but are not listed there.

Although, the categories can be used to determine similarity between pages, clearly the Wikipedia category system is "noisy". In other words, there are opportunities in the category system for further refinement. On a different note, our technique can help find subsets of pages that are very closely related. Then one can apply category refinement on those pages using the clustering results of our method.

B. Wikipedia Page Links

Each wiki page or article usually has one or more references to other articles. We refer to them as page links. If a link to a page p appears on other wiki articles, we consider them as incoming links to p. On the other hand, the links that appear on a page p is considered outgoing links from p. Figure 2 also shows the excerpts from the both "Facebook" and "Myspace" pages containing an outgoing link to the "Google" wiki article. We consider the fraction of common incoming and outgoing links in two wiki articles as their similarity measure. The detail of the how these quality measures are computed can be found in the next subsection.

C. Cluster Quality Measure

In order to determine the quality of the clusters, we computed the intra-cluster distance of the obtained clusters using our validation criteria. As we mentioned earlier, the validation criteria are the wiki category data and page link data. Please note that these data are independent of the editing

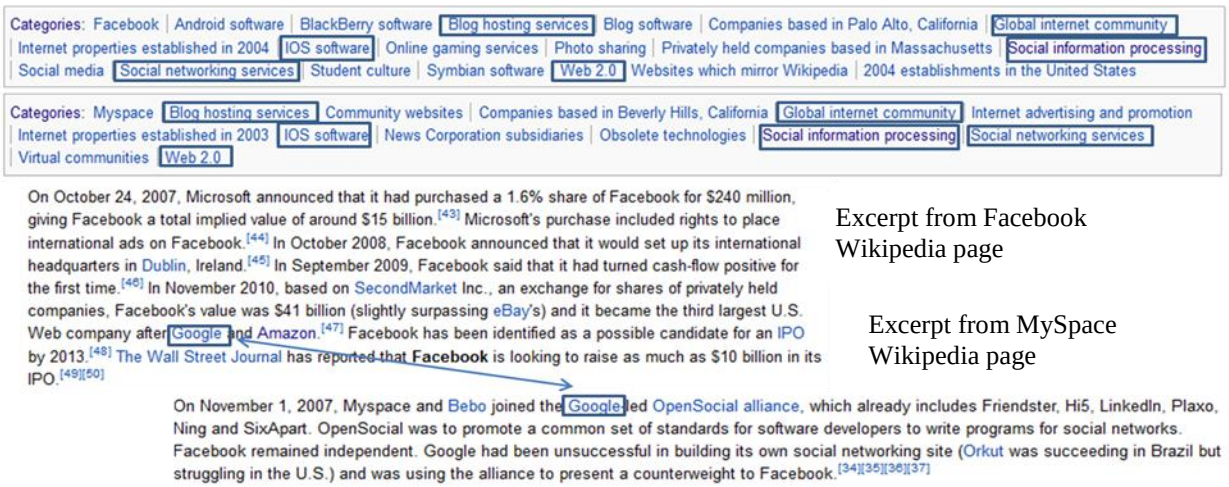

Figure 2. The six common categories from the Facebook and Myspace Wikipedia pages are shown. The excepts from these two pages show a outgoing link to the "Google" article.

dataset we use to obtain the wiki page clusters in the previous step.

The similarity between two Wikipedia pages is determined using Jaccard similarity Index on these two criteria. The Jaccard similarity between two sets of objects a and b can be computed as: $J(a,b) = |a \cap b|/|a \cup b|$. The distance is measured by $1 - J(a,b)$.

Category-based measure: We denote the set of categories for a Wikipedia page $p \epsilon S_l$ as $C(p)$. The distance (based on the categories) between two pages p_1 and p_2 in the same cluster is measured as

$$d^C(p_1,p_2) = 1 - \frac{|C(p_1) \cap C(p_2)|}{|C(p_1) \cup C(p_2)|}$$

The intra-cluster distance of the cluster S_l is defined as:

$$Intra^C(S_l) = \frac{2}{|S_l| \times (|S_l|-1)} \sum_{p_i,p_j \epsilon S_l} d^C(p_i,p_j)$$

Finally, the average intra-cluster distance for K clusters is determined by: $Avg^C(K) = \frac{1}{K}\sum_{l=1}^{K} Intra^C(S_l)$

Page Link-based measure (internal links): We represent the set of the pages that link to a page p in the cluster S_l as $L_{in}(p)$. Let us consider the incoming links to p that are only internal to the cluster S_l, in other words we are considering the pages in S_l, each of which has at least one reference to the page p. We denote this set by $L_{in}(p,S_l) = L_{in}(p) \cap S_l$.

Similarly, for the outgoing links internal to the cluster S_l, we have $\mathrm{L_{out}(p,S_l) = L_{out}(p) \cap S_l}$. Therefore, the distance measure based on the links internal to the cluster S_l is the following:

$$d^{L(internal)}(p_1,p_2,S_l) = 1 - w_1 \times \frac{|L_{in}(p_1,S_l) \cap L_{in}(p_2,S_l)|}{|L_{in}(p_1,S_l) \cup L_{in}(p_2,S_l)|} - w_2 \times \frac{|L_{out}(p_1,S_l) \cap L_{out}(p_2,S_l)|}{|L_{out}(p_1,S_l) \cup L_{out}(p_2,S_l)|}$$

In our experiments we use $w_1 = 0.5$ and $w_2 = 0.5$.

The intra-cluster distance for the cluster S_l and the average intra-cluster distance for K are determined in the same way we did for the category based measure:

$$Intra^{L(internal)}(S_l) = \frac{2}{|S_l| \times (|S_l|-1)} \sum_{p_i,p_j \epsilon S_l} d^{L(internal)}(p_i,p_j,S_l)$$

$$Avg^{L(internal)}(K) = \frac{1}{K}\sum_{l=1}^{K} Intra^{L(internal)}(S_l)$$

Page Link based measure (external links): Now let us consider the pages external to the cluster S_l and have links to a page $p \epsilon S_l$. We denote this set of external incoming links by $\mathrm{L_{in}(p,\tilde{S}_l) = L_{in}(p) \backslash S_l}$. Similarly, the outgoing links external to the cluster $S_l$would be: $\mathrm{L_{out}(p,\tilde{S}_l) = L_{out}(p) \backslash S_l}$.

We obtain distance between the pages $\mathrm{p_1, p_2 \epsilon S_l}$:

$$\mathrm{d^{L(external)}(p_1,p_2,S_l) = 1 - w_1 \times \frac{|L_{in}(p_1,\tilde{S}_l) \cap L_{in}(p_2,\tilde{S}_l)|}{|L_{in}(p_1,\tilde{S}_l) \cup L_{in}(p_2,\tilde{S}_l)|} - w_2 \times \frac{|L_{out}(p_1,\tilde{S}_l) \cap L_{out}(p_2,\tilde{S}_l)|}{|L_{out}(p_1,\tilde{S}_l) \cup L_{out}(p_2,\tilde{S}_l)|}}$$

where $w_1 = 0.5$ and $w_2 = 0.5$.

The intra-cluster distance for the cluster S_l and the average intra-cluster distance for K are as follows:

$$Intra^{L(external)}(S_l) = \frac{2}{|S_l| \times (|S_l|-1)} \sum_{p_i,p_j \epsilon S_l} d^{L(external)}(p_i,p_j,S_l)$$

$$Avg^{L(external)}(K) = \frac{1}{K}\sum_{l=1}^{K} Intra^{L(external)}(S_l)$$

Also note that $L_{in}(p) = L_{in}(p,S_l) \cup L_{in}(p,\tilde{S}_l)$ and $L_{out}(p) = L_{out}(p,S_l) \cup L_{out}(p,\tilde{S}_l)$.

In our experiments, to compute the average intra-cluster distance we only considered clusters with the size larger than 1

or $|S_l| > 1$. Therefore, the average intra-cluster distance based on categories now becomes:

$$Avg^C(K) = \frac{1}{K'}\sum_{l=1}^{K'} Intra^C(S_l)$$

where, $K' = |\{S_l: \ |S_l| > 1, \ l = 1, \dots, K\}|$ and $K \geq K'$.

The average intra-cluster distance for other measures are computed similarly.

D. Random Cluster Generation:

We compare our results with the quality measures computed from the random clusters. The randomized clusters are generated from the "editor behaviors" based on clustering results. We have adopted two different approaches for randomization.

Random clusters with the same cluster size distribution: In our first approach, we generate random clusters for a specific *K* by shuffling the Wiki pages across different clusters obtained from *k*-means (based on "editor behaviors"). We keep the same cluster size distribution as the original clusters. We compute $\text{Avg}^{\text{C}}(\text{K})$ from the random clusters using the using the Wikipedia category data and compare the results with that of the original clustering.

Randomized page link graph: We can obtain a directed page link graph $G_{PL} = (V_{PL}, E_{PL})$ from the Wikipedia page links. Here, V_{PL} is the set of all pages being considered in the clustering and E_{PL} is the set of incoming and outgoing links of the pages. In our second approach, we actually do not generate random clusters. Rather we randomize the page link graph, G_{PL} and generate $R_{PL}(d_1, \dots, d_{|V_{PL}|}; o_1, \dots, o_{|V_{PL}|})$, where the degree in-degree $(d_1, \dots, d_{|V_{PL}|})$ and out-degree distribution $(o_1, \dots, o_{|V_{PL}|})$ of each node in the original graph, G_{PL} are preserved.

Then we compute the average intra-cluster distances based on both the internal and external links ($Avg^{L(internal)}(K)$ and $Avg^{L(external)}(K)$) based on R_{PL} and compare those with results obtained using G_{PL}.

In our experiments, we considered multiple iterations to generate the random clusters and computed the cluster quality measures.

θ	2001	2002	2003	2004	2005
3	0.5357	0.6055	0.6129	0.533	0.4702
4	0.5809	0.6081	0.623	0.53	0.4739
5	0.58	0.607	0.6314	0.529	0.4679

Table 1: Average clustering coefficient

V. RESULTS

As we mentioned earlier, the Wikipedia data set used in our experiments consisted of the edits that have taken place between 2001 and 2008. We extracted the collaboration graphs from the edit data with the threshold values, $\theta = 3, 4$ and 5 and considered one snapshot for every year. Table 1 lists examples of the clustering coefficients from the years 2001 through 2005. It is observed that the clustering coefficient is very high for the collaboration graphs. We considered the collaboration graphs with $\theta = 5$ in our analysis.

First, we performed *k*-means clustering based on the individual editor behavior. We pick 8000 most active pages and 5000 most active editors. Then we computed the measures $\text{Avg}^{\text{C}}(\text{K})$, $\text{Avg}^{\text{L(internal)}}(\text{K})$ and $\text{Avg}^{\text{L(external)}}(\text{K})$ for different *K* values based on the category and page link data we described earlier. Then we generated the random clusters and computed the average intra-cluster distance in order to compare with the actual clusters.

Next, we generated the *k*-means clusters based on the community-based editing behavior described in the section III. Some of the key results from the individual editor behaviors are depicted in the Figures 3, 4, 5 and 6.

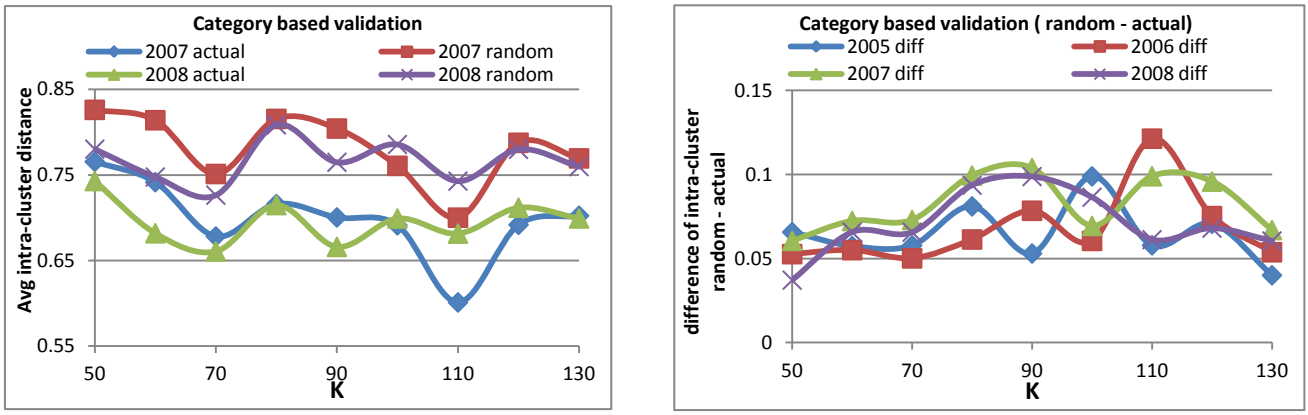

Figure 3. On the left: Comparison of the actual and random intra-cluster distances (individual editor behavior) during 2007-2008 based on the category based data. On the right: Positive difference between the random and acutal intra-cluster distances (individual editor behavior) during 2005-2008.

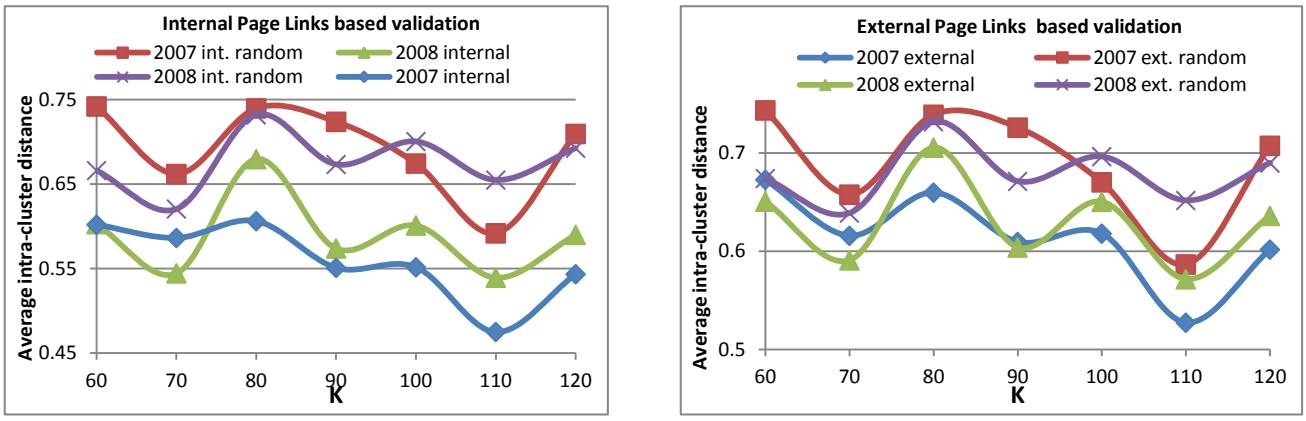

Figure 4. On the left: Comparison of the actual and random intra-cluster distances(individual editor behavior) from 2007 and 2008 based on the page links internal to the clusters. On the right: Comparsion of the actual and the random intra-cluster distances from 2007 and 2008 based on the external page links

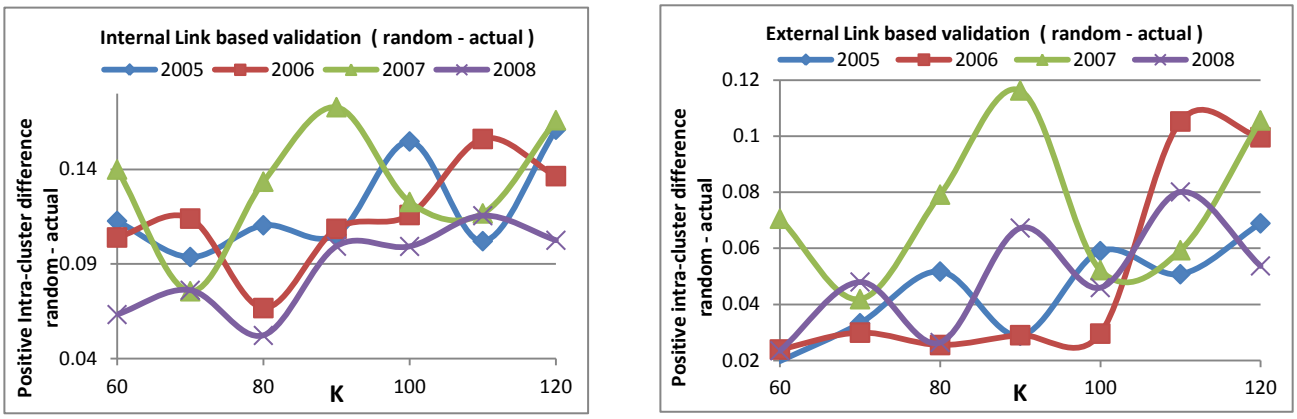

Figure 5. On the left: Difference of the random and actual intra-cluster distances (individual editor behavior) during 2005-2008 based on the page links internal to the clusters. On the right: Difference of the distances during 2005-2008 (individual editor behavior) based on the external page links.

We first consider the results from the individual editor behaviors. The left plot of Figure 3 shows the comparison of the $Avg^{C}(K)$ from both the actual and random clusters obtained in 2007 and 2008 for the *K* values 60-120 with step size 10. The results show that the actual distance is always offset by approximately 0.05 - 0.1 from the random distance (5 to 10% smaller than random). The result is more evident from the right plot where we observe the difference of the actual and random distances during the years 2005-2008. The mean difference is roughly 7% which is significant given that the random clusters follow the exact size distribution and very similar trend of the actual clusters. The peak difference is observed around K = 90-110. Therefore, we speculate that the number of major topics/categories of the Wikipedia pages would be closer to 100.

We computed $Avg^{L(internal)}(K)$ and $Avg^{L(external)}(K)$ based on the page links next. Figure 4 shows the results from the years 2007 and 2008 based on both the internal and external page links. We found very similar results as in the case of the category-based validation. However, both the actual and random distances are now smaller than the category based validation results. This reflects the inherent "noisiness" of the Wikipedia categories.

Figure 5 depicts that for higher values of K ($K \geq 90$) the average difference between the random and actual distances is around 0.1 (10% difference) in the case of internal links. However, for the external link based validation, the average difference is slightly lower (around 6% difference for $K \geq 100$) . The average difference is noticeably rising for even higher values of K ($K \geq 120$) for both internal and external page links.

Finally, Figure 6 depicts some of the category validation results from the community based *k*-means clustering. As we described in section III, we identified overlapping communities through the connected iterative scan (CIS) algorithm. We conjectured that an editor may have multiple interests and belong to different communities. Therefore, instead of individual edits we considered edits by the overlapping communities and obtained the clustering of the wiki pages based on community editor behavior. We validated the results from the years 2004 and 2005 using category data. We found that the results are actually very close to random. The reason

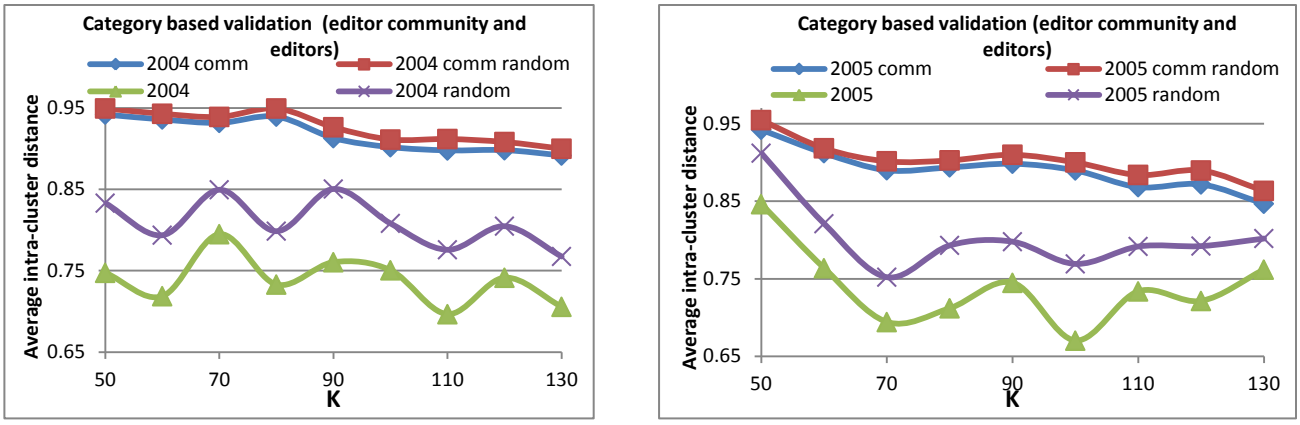

Figure 6. On the left: Comparison of the intra-cluster distances from the actual and random clusters (both individual and community editor behavior) from 2004 based on the category based data. On the right: Comparison of the distances from 2005 (both individual and community) based on category data.

primarily is that there exists a very large component among the communities which includes almost every editor. When we consider the community behavior, almost every page is edited by the largest component and the homophily of individual editor behavior is destroyed. Another interpretation of this result would be that there are perhaps not many editors who actually have overlapping interests. The results from the individual editor behavior reflect this fact.

VI. Related Work

In this section, we discuss related work that study Wikipedia data, and its evolution, categories, classifications and collaboration networks.

A. Evolution and growth of pages

Buriol et al. analyzed the evolution of Wikipedia as a dynamic graph using pages as nodes with temporal stamps and page links as edges [5]. The overall density of pages and links is increasing and different aspects of the Wikigraph have reached distinct stages of evolution. Capocci et al. studied the growth of Wikipedia page network in [4]. In their work, local rules, for example preferential attachment, are observed during adding new pages, while editors act globally on the network to steer its evolution. Wilkinson and Huberman explained the number of edits to pages using a stochastic mechanism [6]. The most popular pages are most visible and likely to be edited more. The quality of pages is also positively correlated with the number of edits. Finally, Das and Magdon-Ismail demonstrate information growth of Wikipedia pages as a collective wisdom [8]. Their model reflects the natural features of Wikipedia edits, including not only the rise of edits as a page becomes popular, but also a decay of edit rate as the page becomes nearly perfect.

B. Categories and classification of pages

Classification of the Wikipedia pages can be achieved through network analysis of their page links. Compared to original top-down categories of Wikipedia, clustering methods, which perform partition of the page network based on page links, result in significant different clusters. This result suggests that page-link in Wikipedia is not a direct indicator for similarity [9]. Holloway et al. analyzed the semantic structure of Wikipedia pages to show that categories associated with individual pages display a power-law distribution, and can be used to measure category similarity [10]. In this case, Wikipedia categories appear to be well-clustered and well-maintained.

C. Edit network

Information of pages and edit behavior are closely correlated. The relationship and interaction between pages and editors are analyzed by Jesus, et al. with a bipartite network of pages and editors [11]. Small subsets of topics are selected. Pages that are edited by common authors form densely connected cliques. These cliques are further clustered into larger connected modules, which reflect a broader common interest of editors. Brandes et al. show that their mathematic models that analyze collaboration of editors can help identify certain types of editors [12]. In addition, several quantitative indicators are proposed to reflect global structure of edit network, and to indicate quality of Wikipedia pages in [12]. Iba et al. investigated edit behavior as a dynamic social network to specify the most creative editors, who start and contribute high quality pages, from the most active editors [13]. These editors are associated to two major categories of pages: focused areas edited by a few experts and broad topics edited by a large number of editors.

VII. Discussions

We observed the lifetime and the edit pattern of 163 most popular pages on the Wikipedia for the years 2004 and 2005. We tracked edit behavior of these pages from 2001 to 2007 and calculated their average number of edits in each quarter. The numbers show that these pages are most heavily edited in continuous time windows, which are about 4 quarters of a year. This observation is consistent with a previous report that edit rates of the most heavily edited pages drop by nearly half after 400 days [8]. Therefore, a collection of edit behavior to

pages in a one-year window would be able to cover most edits to certain pages, whose edit peaks fall in one calendar year. We selected the most heavily edited pages from snapshots from one year.

Another example of collective information is blog, where decentralized classification instead of hierarchical classification is applied. This is because hierarchical classification takes more labor and gives less freedom to users. Therefore, tagging has been used to organize and cluster blogs, in which classification heavily relies on choices of tags by users [14]. Tags can be duplicated or lack of shared meaning. Clustering of blogs using tags usually relies on the removal of non-English tags, while information of editing behavior will be able to handle both English and non-English blogs.

VIII. Conclusions and Future Directions

In this paper, we investigate whether the editor history of the Wikipedia pages can alone be used to identify article groups that are very closely related. We used both individual editor and overlapping community based edit behaviors to perform clustering on the pages. Then we examined the "meaningfulness" or "compactness" of the clusters in the validation phase using Wikipedia category and page links data. We compared the result with the random clusters and observed that the homophily of the individual editor behavior is reflected onto the groups of pages with similar topics. On the other hand, we observed that the overlapping community based edit behavior did not yield very meaningful clusters. Perhaps not many editors actually have overlapping interests in topics.

We also studied page edit dynamics which can be potentially incorporated into the clustering of the pages to achieve better quality of clusters. We can also use Wikipedia category hierarchy to determine similarity of the pages. It will also be worthwhile to look at the traditional cluster quality measures [16][17][18][19] to validate our findings.

Finally, the editor behavior based clustering of pages can help refine the Wikipedia category system. It can also be useful in situations such as finding similar blog posts from the web by considering the edits and comments by the individuals.

References

[1] D. Easley and J. Kleinberg, "Networks, Crowds, and Markets - Reasoning About a Highly Connected World", 1st ed., Cambrige University Press, 2010.

[2] Wikipedia categories: http://en.wikipedia.org/wiki/Wikipedia:Categorization#Categories_do_not_form_a_tree

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

[4] Capocci A, Servedio VD, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G. "Preferential attachment in the growth of social networks: the internet encyclopedia Wikipedia.", Phys Rev E Stat Nonlin Soft Matter Phys. Sep 2006.

[5] Luciana S. Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. "Temporal Analysis of the Wikigraph", In Proc. of Web Intelligence, page 45-51, Hong Kong, 2006.

[6] Wilkinson, D. M. and Huberman, B. A. "Assessing the value of coooperation in Wikipedia". First Monday 12, 4. 2007.

[7] M.Goldberg, M. Krishnamoorthy, M. M-Ismail and N. Preston, "Clustering communities by clustering a graph into overlapping subgraphs", Proceedingsof IADAIS conference on applied computing, 2005, pp. 97-104.

[8] S. Das and M. M-Ismail, "A Model for Information Growth in Collective Wisdom Processes", ACM Transactions on Knowledge Discovery from Data (TKDD), 2011.

[9] A Capocci, F Rao, G Caldarelli, "Taxonomy and clustering in collaborative systems: the case of the on-line encyclopedia Wikipedia", Europhysics Letters (2007), Volume: 81, Issue: 2, Pages: 5

[10] Todd Holloway, Miran Bozicevic, and Katy Borner. 2007. "Analyzing and visualizing the semantic coverage of Wikipedia and its authors". Research Articles. Complex. 12, 3 (January 2007), 30-40.

[11] Rut Jesus, Martin Schwartz, and Sune Lehmann. 2009. "Bipartite networks of Wikipedia's articles and authors: a meso-level approach." In Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09). ACM, New York, NY, USA, , Article 5.

[12] Ulrik Brandes, Patrick Kenis, Jurgen Lerner, and Denise van Raaij. 2009. "Network analysis of collaboration structure in Wikipedia". In Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, NY, USA, 731-740.

[13] Takashi Iba, Keiichi Nemoto, Bernd Peters, Peter A Gloor. "Analyzing the Creative Editing Behavior of Wikipedia EditorsThrough Dynamic Social Network Analysis", Procedia Social and Behavioral Sciences (2010) Volume: 2, Issue: 4, Pages: 6441-6456.

[14] Christopher H. Brooks and Nancy Montanez. 2006. "Improved annotation of the blogosphere via autotagging and hierarchical clustering". In Proceedings of the 15th international conference on World Wide Web (WWW '06). ACM.

[15] S. Kelley, M. Goldberg, M. M-Ismail, K. Mertsalov and W. Wallace, "Defining and Discovering Communities in Social Networks", in *Handbook of Optimization in Complex Networks*, eds. M. Thai and P. Pardalos, Chapter VI, 2011.

[16] Davies, D.L., Bouldin, D.W. (2000) A cluster separation measure. IEEE Trans.Pattern Anal. Machine Intell., 1(4), 224-227.

[17] Dunn, J. (1974) Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics ,4, 95-104.

[18] Hubert, L. and Schultz, J. (1976) Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychology, 29, 190-241.

[19] Rousseeuw, P.J., (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.