Classifying Wikipedia Articles Using Network Motif Counts and Ratios

Guangyu Wu
School of Computer Science
and Informatics
University College Dublin
Dublin, Ireland
guangyu.wu@ucd.ie

Martin Harrigan
School of Computer Science
and Informatics
University College Dublin
Dublin, Ireland
martin.harrigan@ucd.ie

Pádraig Cunningham
School of Computer Science
and Informatics
University College Dublin
Dublin, Ireland
padraig.cunningham@ucd.ie

ABSTRACT

Because the production of Wikipedia articles is a collaborative process, the edit network around a article can tell us something about the quality of that article. Articles that have received little attention will have sparse networks; at the other end of the spectrum, articles that are Wikipedia battle grounds will have very crowded networks. In this paper we evaluate the idea of characterizing edit networks as a vector of motif counts that can be used in clustering and classification. Our objective is not immediately to develop a powerful classifier but to assess what is the *signal* in network motifs. We show that this motif count vector representation is effective for classifying articles on the Wikipedia quality scale. We further show that ratios of motif counts can effectively overcome normalization problems when comparing networks of radically different sizes.

Categories and Subject Descriptors

E.0 [**Data**]: General – Data quality; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms

Keywords

Wikipedia Quality, Edit Networks

1. INTRODUCTION

A key principle in network data analysis is that the network structure around a node can tell us something about the characteristics of that node. This fundamental idea has been demonstrated in areas as diverse as spam filtering [5, 4], telecommunications [3], bioinformatics [20] and social network analysis [13]. How best to represent the network around a node is still a significant research challenge. Recently profiling using network motif counts has emerged as a promising solution for characterizing networks [20, 25, 3]. In this paper we apply this idea to the classification of Wikipedia articles in terms of quality. Figure 1 shows the edit networks around two articles from the Wikipedia History project. The first article is a "Start" class article while the second article is a "Featured Article", the very top of the Wikipedia quality scale. It is clear from an analysis of these networks that there are differences between them. It is perhaps surprising that the Start class article has such a low quality score given that it has received input from so many contributors. After all, a key motivation in Internet collaboration is that *many eyes* on a piece of work will result in a good quality product [21] – or at least a low error product. Our preliminary work on Wikipedia quality has shown that this is not necessarily the case [25]. It is not sufficient for an article to have received attention from many contributors, it is important that these contributors are themselves experienced. This experience is evident in the Featured Article in Figure 1 where contributors to the key article have also collaborated on other articles in the network.

We present an assessment of this network motif-based characterization of Wikipedia articles on three datasets gathered from Wikipedia. The evaluation covers over 3,000 articles from the Wikipedia projects on *History, United States* and *Meteorology* (see section 4 for details). In the next section we provide some details on the nature of the edit networks and the design decisions to be made in extracting them. Details of the network motif counting process are presented in section 3. Sections 5 and 6 present some results on classification and the impact of feature selection on classification accuracy. The final section of the paper (section 7) shows how classification accuracy can be improved by looking at count ratios rather than normalized counts.

2. WIKIPEDIA ANALYSIS

In contrast to traditional encyclopedias, where authority derives from expert contributors, Wikipedia depends on a mixture of expertise, collaboration and consensus to produce quality articles. There has been some controversial research that suggests that the quality of Wikipedia articles approaches that of established encyclopedias [10]. The famous quote from Surowiecki's *The Wisdom of Crowds* is that "under the right circumstances, groups are remarkably intelligent" [22].

The collaborative and open nature of Wikipedia makes it very receptive to the *many eyes* idea [16] – a large number

WikiSym '12, Aug 27-29, 2012, Linz, Austria.
Copyright © 2012 ACM 978-1-4503-1605-7/12/08 ... $15.00

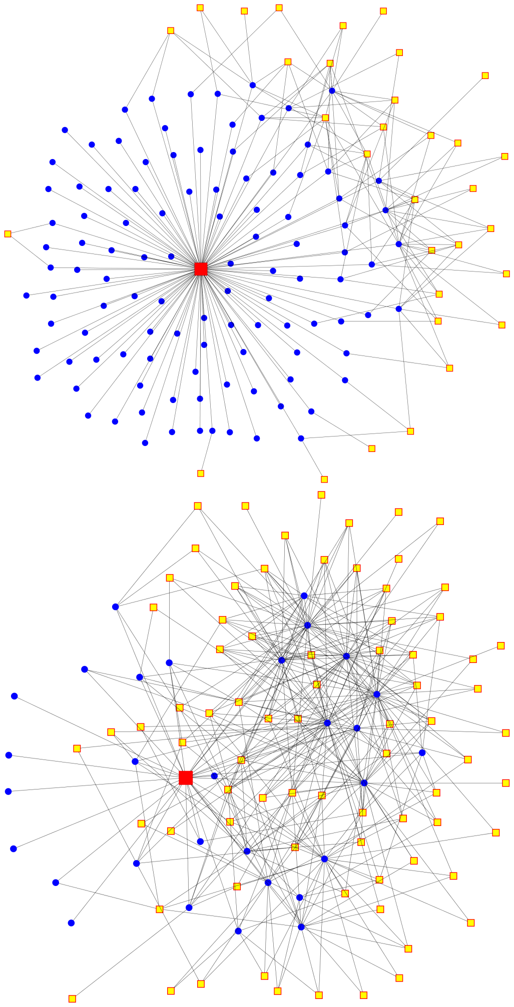

Figure 1: Two sample edit networks from Wikipedia. The red squares are the key articles, the yellow squares are other articles and the blue circles are editors. The top article is "Family History" a Start Class article, and the bottom article is "Eardwulf of Northumbria" a Featured Article.

of contributors can cooperate to produce a quality article. However, the fact that the upper network in Figure 1 is a Start class article suggests that this is not sufficient. It is also important that this collaboration has been constructive and it is better if the editors have a reasonable reputation as contributors. Adler and De Alfaro [1] have pursued a content driven strategy to assess editor reputation. They have used text survival and edit distance to quantify editor reputation. In later work [2] they show that edit longevity is a good measure of editor contribution.

Korfiatis et al. [14] pursue a network-based strategy to evaluate authoritative sources in Wikipedia. They construct a two-mode network of articles and contributors. The article nodes are linked by hyperlinks and contributors are linked if they have worked on the same article. Contributors are also linked to articles on which they worked. The study proposes article and contributor degree centrality as indicators of authoritativeness. This is similar in spirit to the strategy in our work as degree centrality is captured by a subset of the network motifs we consider.

Brandes et al. [6] have also analysed the collaboration structure in Wikipedia. Their work has focused on the edit interactions on individual articles. Edges between individual contributors represent *delete*, *undelete* and *restore* interactions. The main contribution of this work is to present the notion of bipolarity that captures the level of conflict between the contributors to an article. Thus the work is more directed at the problem of Wikipedia vandalism than the issue of authoritativeness that is the subject of this paper.

Recently, Laniado et al. [15] presented an algorithm that assigns scores to all contributors of a Wikipedia article according to their contribution, and selects the top contributors to build a collaboration network of authors where edges represent the co-authorship between authors. Thus the inexperienced authors are filtered out and the co-authorship networks become more informative. With the exception of eigenvector centrality (where edge weights were considered) the features they extracted were taken from unweighted versions of the networks.

Dalip et al. [8] presented a comprehensive assessment of quality indicators in collaborative content curation with a focus on Wikipedia. In their analysis they considered 69 indicators including text features, review features and basic network features. They used a machine learning approach to discover the most effective indicators and combination of indicators. They found that the easy-to-extract text-based features were most informative – more informative that more complex features based on link analysis.

There also exists non-network based studies, for example, Lipka et al. [17] used machine learning techniques to identify featured articles using character trigram and part-of-speech trigram vectors. These features that are known to be characteristic of writing style out-performed alternatives in both a single domain and a domain transfer situation with F-measure scores of 0.88 across domains and good performance on articles of varying length. The work of Javanmardi et al. [12] on vandalism detection is in the same spirit as this. They show that a content-based strategy can be very effective for identifying vandalism edits in Wikipedia.

The work by Dalip et al. [8] and Lipka et al. [17] is complementary to ours in that our network-based features can be combined with their content-based features to further improve classification accuracy.

2.1 Wikipedia Networks

Our analysis is carried out on networks of the type shown in Figure 1. These networks are constructed from the edit histories of the articles that can be retrieved from Wikipedia. To build our edit networks we applied several rules to filter the raw data. Firstly, as some articles have a long edit history, we only considered the last 200 revisions of the articles, and extracted the editors who made these revisions. We also considered all articles that are connected by hyperlinks from the originating or ego articles. We retained the linked-to articles that have been edited by at least one of the ego article editors – it is easy to see this in the top network in Figure 1. Editors often repeatedly save their changes during a short session, so we judged continuous revisions by the same editor as a single revision.

In Wikipedia there are two type of editors, registered and unregistered users, where unregistered users are automatically named by the IP addresses they used when editing. We drop the unregistered users from the network as they don't result in interesting network structure.

Bots are allowed by Wikipedia to do some automatic editing and conventionally use names starting or ending with 'bot'. Bots perform a huge amount of small editing tasks so the bots are often very high-degree nodes in the network. For this reason we drop the bots from the network as their high-degree distorts our network motif counting results. Furthermore we do not expect that the level of attention from bots should impact on the quality of the article.

In constructing these ego-networks it has been important to catch any reorganization such as article renaming or merging. In these cases, Wikipedia automatically redirects old articles to newer ones. Before building article ego-networks, we tested all the articles in order to identify any newer destination articles. The revision articles we used to extract article edit history for ego-networks are based on the destination articles in our experiments.

3. MOTIFS AND MOTIF COUNTING

This section describes the motif profile we set up for the edit networks and the process of acquiring motif counts for each network.

3.1 Wikipedia Network Motifs

Our Wikipedia network motifs comprise editor and article nodes and editor-article edges (see Figure 2). The editor-article edges represent edit activities on Wikipedia articles. The networks are bipartite since there are no between-editor edges or hyperlink edges between articles. Hyperlink edges were excluded from our consideration because earlier analysis [24] has found that hyperlink density can dominate the network motif profiles, and from a quality perspective, this is not an interesting distinction between articles.

The complete set of network motifs of up to five nodes is shown in Figure 2. In this figure, we organize the motifs in a tree structure that will be discussed in Section 7.

We used **nauty** [18] to enumerate all network motifs up to five nodes without considering node labels. There are 31 unlabeled network motifs with between one and five nodes. When we allow nodes to be either 'editor' or 'article' these 31 unlabeled network motifs produce 419 two-labeled network motifs. When motifs with nodes of just one type and motifs with editor-editor or article-article edges are removed the set reduces to the 17 network motifs in Figure 2. The motif hierarchy has four layers where each layer contains motifs according to the number of nodes in them.

3.2 Motif Counting

Before presenting the details of our motif counting process, it is important to mention that there are two different definitions of a network motif that can be applied, the motif can be *induced* or *non-induced*.

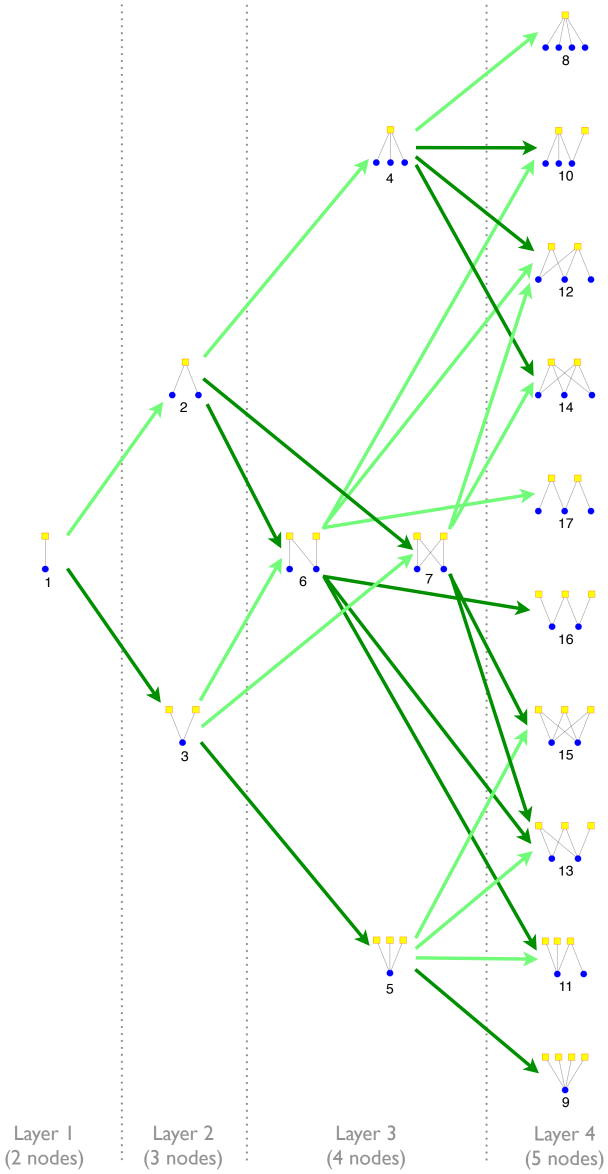

Figure 2: Motif Hierarchy: 17 motifs up to five nodes where yellow squares are articles and blue circles are editors. Dark green arrows indicate the addition of an article and light green arrows indicate the addition of an editor.

An induced motif must contain all edges between its nodes that are present in the target network, whereas a non-induced motif need not. Induced motifs are a subset of the non-induced motifs. Figure 3 shows an example that explains the difference. There are only two induced instances of the motif shown in (a) in the network (b), these are 1-3-4 and 2-3-4. 1-2-3 is not an induced instance because of the edge 1-3. Counting non-induced motifs returns five instances (1-2-3, 2-3-1, 3-2-1, 1-3-4 and 2-3-4).

Clearly, counts of non-induced motifs will be greater than

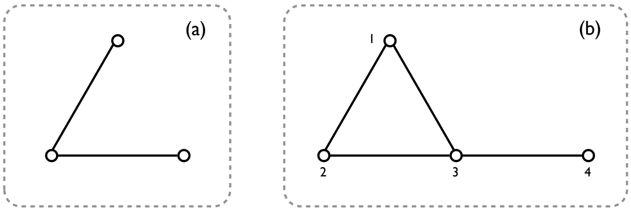

Figure 3: Induced Subgraph Counting: only two induced subgraph instances (1-3-4 and 2-3-4) exist in *Graph (b)* for *Subgraph (a)*.

counts of induced motifs. However, the counts are related [9] and our own evaluation has shown that the two alternatives result in very similar classification accuracies. Pržulj [20] has also shown that the two variants lead to similar results. We use induced motif counts in this analysis.

We use `FANMOD` [23] to obtain the number of motif instances in an ego-network. `FANMOD` is designed to output all induced subgraph instances for a particular size in a given target network. We ran `FANMOD` on each article ego-network to produce the motif counts comprising 3, 4 and 5 nodes. In addition, the counts for the 2-node motif is calculated as it is equal to the number of edges in the ego-network.

Thus each article is represented as a vector of motif counts of length 17. The classification analysis discussed in the next section is performed on this un-normalized data – we experimented with L2 normalization [11] but it did not improve accuracy.

4. DATASETS

The experiments were based on the articles from three collections on *History*, *United States* and *Meteorology.* These collections were selected because they include a large number of articles, especially a sufficient number of *Featured Articles.* The official descriptions for the different classes in the Wikipedia quality scale are shown in Table 1[1].

Class	Summary
Featured Article	“Professional, outstanding, and thorough; a definitive source for encyclopedic information.”
A	“A fairly complete treatment of the subject.”
Good Article	“Useful to nearly all readers, with no obvious problems...”
B	“The article is mostly complete and without major issues...”
C	“The article is substantial, but is still missing important content or contains a lot of irrelevant material.”
Start	“An article that is developing, but which is quite incomplete and may require further reliable sources.”
Stub	“A very basic description of the topic.”

Table 1: Wikipedia quality classes

	Good		Medium	
Dataset	Classes	Count	Classes	Count
History-F-S	F	149	S	299
US-F-S	F	272	S	300
Meteorology-F-S	F	131	S	300
History-FG-CS	FG	440	CS	588
US-FG-CS	FG	565	CS	598
Meteorology-FG-CS	FG	431	CS	600

Table 2: Datasets analyzed

The evaluation considered Wikipedia articles from four different quality classes, Featured articles (F), Good article (G), C-class articles (C) and Start articles (S). We consider the first two classes to be articles of high quality while the last two are of medium or low quality. It is important to state that Start class articles are reasonable sources of information. The really basic articles in Wikipedia are Stub articles. We don’t consider these in the evaluation.

From these collections we created 6 datasets (see Table 2) representing two types of classification challenge, an easy challenge comparing F and S articles and a harder challenge with F and G as the good quality class and C and S as the low/medium class.

The limiting factor in building these datasets is the number of Featured articles available (see Table 2). As the number of articles in other classes greatly exceeded the number of Featured articles, we subsampled the other classes (selecting 300 at random) to ensure that the training data was not too imbalanced.

5. INITIAL CLASSIFICATION

For the classification analysis we consider four methods: random forest (100 trees); logistic regression; k-nearest-neighbor (k-NN) and support vector machine (SVM) [7]. We report performance from 10-fold cross validation tests in terms of overall accuracy and ROC area – ROC area is relevant because overall accuracy may be misleading when errors between classes are imbalanced. Random forest is included because it is an ensemble method that can be expected to give very good performance. Logistic regression is included because it is a simple method that should also perform well and offers some insight into how features contribute to the classification. We include k-NN because the classes may be diverse and a local learner may be expected to work well in these circumstances. SVM is considered because it is a state-of-the-art method that should give good accuracy.

We applied the four classifiers on the three simple datasets (F versus S). The results are presented in Table 3. The best classification accuracy is achieved with logistic regression achieving accuracies above 80% in all cases. This is perhaps not surprising given that we are operating in a 17-dimension space where a linear classifier can produce reasonable accuracy.

A clear pattern in these results is that random forest and logistic regression are performing better than k-NN and SVM.

[1] Details on the Wikipedia Quality Scale are available at: `http://bit.ly/1avQfU`.

	Random Forest		Logistic		*k*-NN		SVM	
Dataset	Accuracy	ROC Area	Accuracy	ROC Area	Accuracy	ROC Area	Accuracy	ROC Area
History-F-S	**82.6%**	**0.89**	80.4%	0.88	76.8%	0.73	79.5%	0.73
US-F-S	85.8%	0.93	**87.2%**	**0.94**	85.0%	0.85	83.4%	0.83
Meteorology-F-S	78.9%	**0.88**	**81.0%**	0.87	75.9%	0.72	75.2%	0.61

Table 3: Classification results for all subgraph instances (F vs. S)

This pattern was maintained in our other evaluations so, to simplify the picture, we do not report further results using *k*-NN or SVM.

In the next evaluation we tackle the larger datasets where high quality articles include both Featured and Good articles and lower quality articles are both C class and Start articles (Table 4). Each of the datasets contains over 1,000 articles (see Table 2) and the classification is more difficult because the distinction between the two classes is less clear. Accuracy falls as expected but roughly two thirds of articles are still classified correctly.

	Random Forest		Logistic	
Dataset	Acc.	ROC Area	Acc.	ROC Area
History-FG-CS	65.3%	0.70	**65.7%**	0.70
USA-FG-CS	70.8%	0.79	**71.8%**	0.79
Meteo-FG-CS	**66.4%**	0.72	60.9%	0.67

Table 4: Classification results for all subgraph instances (FG vs. CS)

6. FEATURE SELECTION

Given our objective of identifying the useful *signal* in edit network motifs for predicting Wikipedia article quality we now turn to the correlations and contributions of individual motifs. If we can identify a small selection of motifs that have the classification power of the full set of motifs then this tells us which motifs are characteristic of good quality collaboration. This will also allow us to simplify the motif characterization process.

6.1 Hierarchical Heatmaps

Our objective here is to cluster motifs based on correlated counts and then select a subset of motifs that are representative of the clusters. When we do this with the History data we get the hierarchical heatmap shown in Figure 4.

If we split the motif set into four clusters according to the obvious sub-trees in the hierarchy the four motif clusters are {M2, M8, M4}, {M9, M15, M5}, {M16, M3, M7, M13, M11} and {M14, M1, M10, M12, M17, M6}.

Next, we would like to select representatives from each of these clusters that are easier to count. While the general problem of network motif counting, addressed by FANMOD [23], is computationally expensive there are certain motifs such as stars and cycles [19] that are easier to count. It transpires that we can select motifs M1, M7, M8 and M9 as representatives of each of the four clusters that satisfy this criterion (highlighted in Figure 4).

6.2 Four Motifs

Classification performance using just this set of four motifs is shown in Figure 5. The first column in each set shows performance with the full set of motifs, the second column in each set shows performance when only four motifs are used. Results with just four motifs are slightly worse by a few percent on average. The worst performance is a drop of 3% when using logistic regression on the Meteorology dataset. There is a slight improvement when using logistic regression on the History dataset. This shows that a characterization of the edit networks using four motif counts captures a lot of information about the quality of Wikipedia articles – it is worth looking in more detail at what these motifs signify.

- M1 is simply a count of the edges in the ego-network. In general, the larger the network the higher will be the quality of the article. That is because a high quality article usually has cited a number of other articles via hyperlinks and the article itself has been revised by many editors hundreds of times. For instance, the mean size of a Featured article in History is 736 edges while the Start class networks have 219 edges in average.
- M7 and indeed the other motifs in its cluster are representative of collaboration by more than one editor on a number of articles. These are *virtuous* motifs that are characteristic of good-quality articles. We have selected M7 from this group because it is easy to count since it is a cycle in the network [19].
- M8 is the selected representative of the cluster of star motifs with articles at the centre. This represents the many-eyes idea where many authors have collaborated on an article. The count for this motif can be calculated directly from the degree of an article node. A node of degree d participates in $\binom{d}{4}$ motifs of type M8 (number of combinations of d objects taking 4 at a time). This motif is not particularly characteristic of quality – the Start class article in Figure 1 has a high count for this motif.
- M9 is the equivalent of M8 but with editors at the centre and is similarly easy to count. These motifs are indicative of good-quality articles.

We have found in other work that the editor's experience is more important than the many-eyes idea for an article to achieve high quality [25] and this is reflected in the effectiveness of M9 and M8 in this classification task.

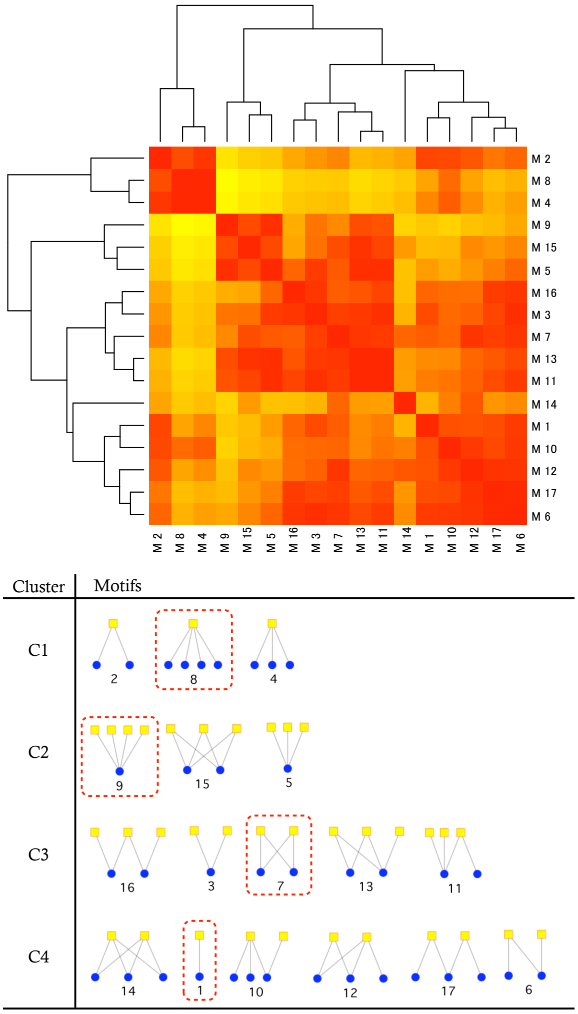

Figure 4: A hierarchical clustering of network motifs. In the heatmap, red (darker) color implies stronger correlation and yellow (lighter) color stands for weaker correlation between motifs.

7. MOTIFS COUNT RATIOS

Representing networks as vectors of motif counts is analogous to the strategy of representing texts as bags of words – and it has all the attendant problems of data normalization. Given that the *relative* abundance of different motifs is the important information, we have examined an alternative representation based on motif ratios.

The motifs can be arranged in a hierarchy as shown in Figure 2 where each child motif contains an extra node and one or more edges. The parent-child relationships in the tree fall into two categories of 13, those involving the addition of an article node (dark green arrows) and those involving the addition of an editor node (light green arrows). This shows for example that M6 instances will be extended from instances of M2 and M3. However, the ratio between counts of M6 and M2 might be different from the ratio between M6 and M3. In the same way, the relative abundance of M7 and M15 should be informative. To explore this alternative representation, we calculated all parent-child ratios in Figure 2 by dividing the child motif count by the parent motif count. When the denominator is zero (parent count) the ratio is set to zero.

Cluster	Ratios
C1	{R5-15, **R3-7**, R7-14, R5-13, R6-12}
C2	{**R6-17**}
C3	{R2-4, R4-8, R7-12, **R6-10**, R1-2, R3-6, R5-11}
C4	{R4-14, R4-12, **R2-7**}
C5	{R6-13, R7-15, R1-3, R4-10, R2-6, R6-16, R7-13, R5-9, **R6-11**, R3-5}

Table 5: A flat clustering of the ratios.

It is worth mentioning that counting induced motifs rather than all motif occurrences raises issues when calculating ratios. For example, while M6 is a sub-network of M7, counts of M6 are not included in M7 because of the 'induced' rule. Figure 10 in the Appendix shows an alternative hierarchy where the ratios are less meaningful for this reason.

7.1 Classification using Motif Count Ratios

Since the motif count ratios are an alternative representation to the motif counts used in the classification evaluation in section 5 we can make a direct comparison between these alternatives.

Figure 5 shows the ratio based results (third columns) compared against all motifs and 4 motifs as discussed already. These results are the best figures obtained in our evaluations. Classification using count ratios is always at least as good as the best alternative. Accuracy when using logistic regression has been improved by 3% on the Meteorology collection, the most difficult of the three datasets. Results are also better on History where the accuracy is up to over 85%. Results are not improved on the USA collection where accuracy is already at 85%. In summary, the use of motif count ratios brings results on all three datasets to 85% or above.

We also assess the effectiveness of ratios on the harder classification task presented in Table 4. These results using logistic regression are shown in Figure 6. The results are consistent, with the ratios showing improvements over the motif counts in all three datasets.

7.2 Ratio Correlations

If we use information gain to identify the ratios that are most predictive for classification we find that the top three ratios are R7-13, R6-11 and R6-16. These all correspond to situations where articles are added to the motifs. These three ratios indicate editor experience and are typical of high quality articles.

Given that the motif count ratios can be strongly correlated in the same way that motif counts are we have prepared hierarchical clusterings of the ratio profiles for each of the three datasets (Figure 7). There is very good correlation between the three hierarchical clusterings. The flat clustering shown in Table 5 shows a partitioning that agrees with all three hierarchies. The clusters in the table are written in the or-

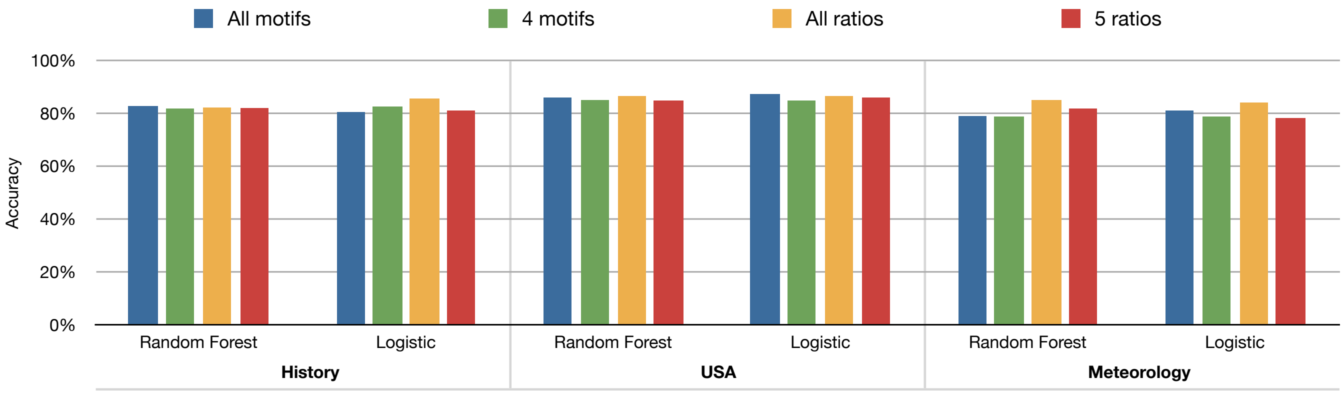

Figure 5: Comparison of classification accuracies when using all motifs, just 4 motifs, all ratios and 5 ratios.

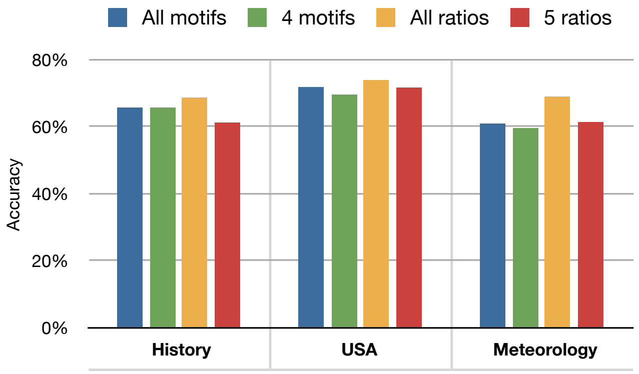

Figure 6: Classification accuracy for logistic regression on the harder classification task (FG-CS).

der taken from the History heatmap (top in Figure 7). The top three clusters (C1, C2, C3) cover ratios that involve the addition of an editor node. The difference between C1 and C2, C3 is that C1 entails the addition of two or three edges while the ratios in C2, C3 add just one edge. On the other side of the hierarchy C4 and C5 entail the addition of an article node. As mentioned at the beginning of this section, the ratios in C5 are the most discriminating.

The next step is to select one ratio from each cluster for classification. We used the ratios that have the largest correlation in total with other ratios in the same cluster. These are effectively the cluster centroids. The five ratios chosen are R2-7, R3-7, R6-10, R6-11 and R6-17 highlighted in Table 5. When we use these five ratios as a proxy for the full set of 26 ratios the fall off in classification accuracy is greater than when we use the subset of four motifs compared with the original motif set (final columns in Figure 5). For instance the performance of logistic regression on the Meteorology collection falls by over 7%.

Again we apply the same analysis on the harder datasets using logistic regression as the classifier. As is the case with the easier datasets, the effect of feature selection on the ratio features is quite damaging – see final column in Figure 6.

In summary, feature subset selection is not as effective with ratios as with motifs. The loss of classification accuracy compared with the full set of ratios is more considerable.

8. MOTIFS, QUALITY AND USER FEEDBACK

Given that the analysis presented here shows how certain collaboration structures are associated with high quality ratings it is worth exploring whether this insight can be used to raise the quality of other pages. At the same time it is worth considering an important alternative 'quality' criterion in Wikipedia. In 2011 an Article Feedback system started to get wide scale deployment in English Wikipedia.[2] This allows readers to rate pages on a five star scale against four criteria, *Trustworthy, Objective, Complete* and *Well-written*. Clearly these scores represent quality criteria that are as important as the quality scores assigned by editors.

In the articles in the three collections we have studied we have 401 articles with 10 or more ratings so we conducted a similar study to see if we could predict feedback ratings from motif profiles. It transpired that the results were surprisingly poor – below 60% in most cases. After looking in to this we found that correlations between Wikipedia quality ratings and user feedback scores were weak. Figure 8 shows this for the Trustworthy criterion on our 401 articles. The results are more or less the same for the other three criteria.

So we seem to have a situation (depicted in Figure 9) whereby collaboration structures, as captured by motif profiling, are predictive of the official Wikipedia quality score but not particularly of user ratings. The really interesting thing here is the third edge in the triangle in Figure 9 showing the poor correlation between quality scores and user ratings. It is clear from Figure 8 that Start class articles are as likely to get five star ratings as are Featured articles. This indicates that the Article Feedback process will be a more useful measure of quality for readers than the official quality scores.

9. CONCLUSIONS

In this paper we explore the hypothesis that an analysis of the edit network around a Wikipedia article provides information about the quality of that article.

[2] http://www.mediawiki.org/wiki/Article_feedback

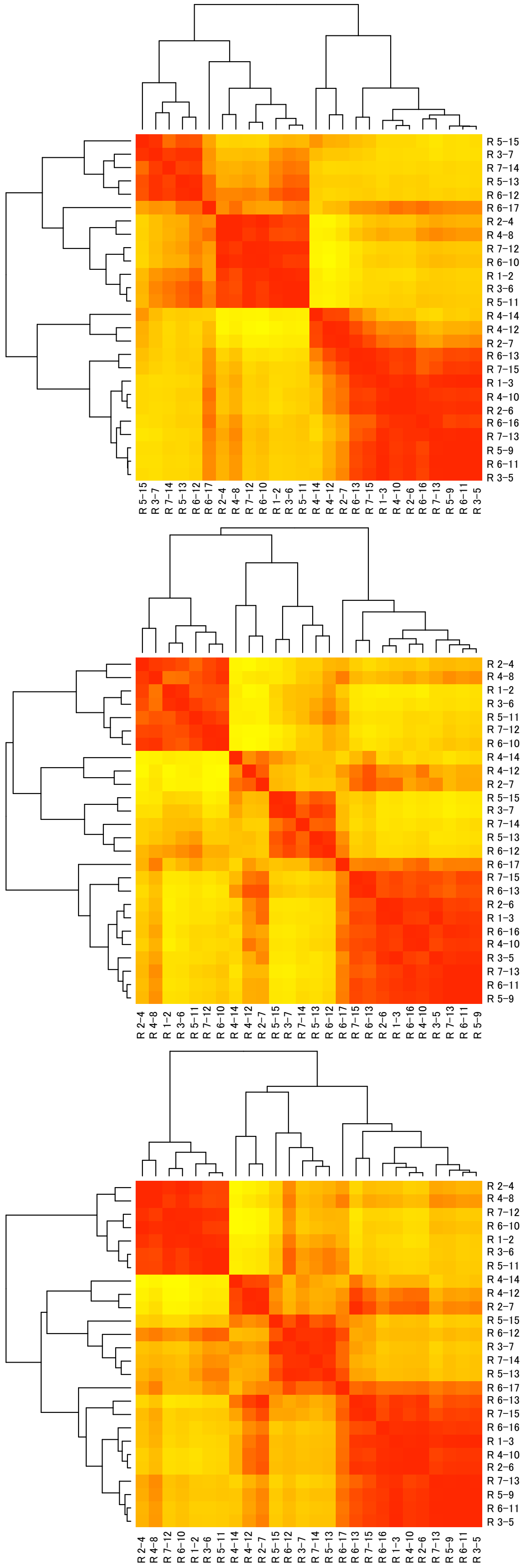

Figure 7: Ratio correlations for History (top), USA and Meteorology.

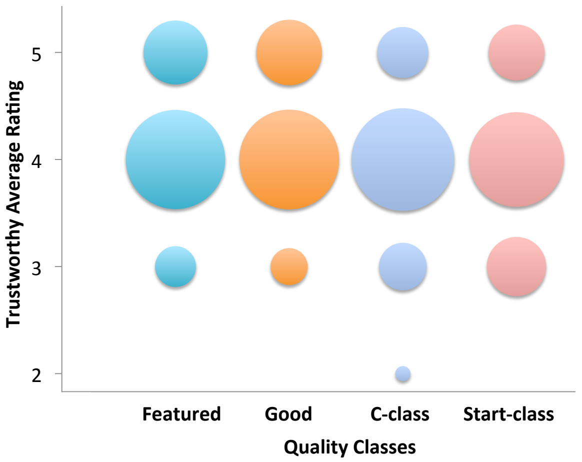

Figure 8: This bubble chart shows article distributions across combinations of feedback and quality scores for the Trustworthy criterion.

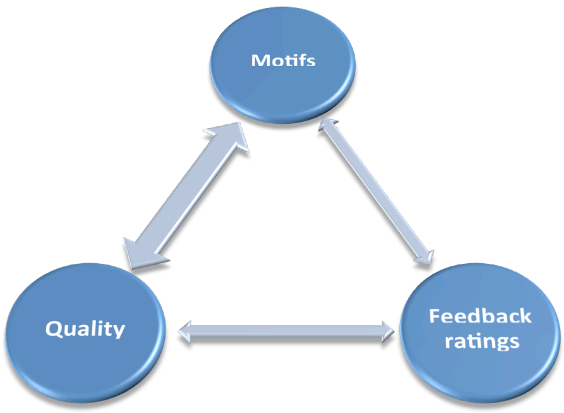

Figure 9: The Wikipedia quality score is correlated with motif profiles but not the Feedback ratings.

We can summarize our main findings about the methodology as follows:

- A feature vector representation based on motif counts is quite predictive of article quality. It can achieve over 80% accuracy on classifying Featured articles against Start class articles and about two thirds accuracy on Featured and Good articles against C and Start class articles.
- The most predictive motifs are those that reflect collaboration with multiple authors collaborating on multiple related articles.
- The motifs are strongly correlated and a subset of just four of the full set of 17 motifs maintains most of the classification power.
- Ratios of motif counts are even more effective than raw motif counts.

The lessons to be learned about Wikipedia quality are more complicated. Pages with good quality scores have characteristic motif profiles, but pages with good user ratings don't. This suggests that a good quality score is evidence that a collaborative curation process has been pursued. However, not all pages with high quality scores get good user ratings and some pages with low quality scores are trusted by users. Perhaps the Wikipedia quality scale is a low error scale rather than a quality scale?

Acknowledgements

This research was supported by Science Foundation Ireland (SFI) Grant No. 08/SRC/I1407: Clique: Graph and Network Analysis Cluster.

10. REFERENCES

[1] B. Adler and L. De Alfaro. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, page 270. ACM, 2007.

[2] B. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the Wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, pages 1–10. ACM, 2008.

[3] E. G. Allan, Jr., W. H. Turkett, Jr., and E. W. Fulp. Using network motifs to identify application protocols. In *Proceedings of the 28th IEEE Conference on Global Telecommunications*, GLOBECOM'09, pages 4266–4272, Piscataway, NJ, USA, 2009. IEEE Press.

[4] L. Becchetti, P. Boldi, C. Castillo, and A. Gionis. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 16–24, New York, NY, USA, 2008. ACM.

[5] P. Boykin and V. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61 – 68, 2005.

[6] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, pages 731–740. ACM, 2009.

[7] M. Cord and P. Cunningham. *Machine learning techniques for multimedia: case studies on organization and retrieval.* Springer-Verlag New York Inc, 2008.

[8] D. Dalip, M. Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by web communities: a case study of Wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 295–304, 2009.

[9] D. Eppstein and E. Spiro. The h-Index of a Graph and its Application to Dynamic Subgraph Statistics. In F. Dehne, M. Gavrilova, J. Sack, and C. Tóth, editors, *Proceedings of the 11th International Symposium on Algorithms and Data Structures (WADS'09)*, pages 278–289. Springer, 2009.

[10] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[11] I. Gradshteĭn, I. Ryzhik, and A. Jeffrey. *Table of integrals, series, and products.* Academic Press, 2000.

[12] S. Javanmardi, D. McDonald, and C. Lopes. Vandalism detection in wikipedia: a high-performing, feature-rich model and its reduction through lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 82–90. ACM, 2011.

[13] K. Juszczyszyn, P. Kazienko, and K. Musiał. Local topology of social network based on motif analysis. In I. Lovrek, R. Howlett, and L. Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 5178 of *Lecture Notes in Computer Science*, pages 97–105. Springer Berlin / Heidelberg, 2008.

[14] N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: an insight from Wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[15] D. Laniado and R. Tasso. Co-authorship 2.0: Patterns of collaboration in Wikipedia. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pages 201–210. ACM, 2011.

[16] A. Lih. Wikipedia as participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource. In *In Proceedings of the 5th International Symposium on Online Journalism*, pages 16–17, 2004.

[17] N. Lipka and B. Stein. Identifying featured articles in Wikipedia: writing style matters. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1147–1148. ACM, 2010.

[18] B. McKay. Practical graph isomorphism. *Congressus Numerantium*, 30(30):47–87, 1981.

[19] K. Paton. An algorithm for finding a fundamental set of cycles of a graph. *Communications of the ACM*, 12(9):514–518, 1969.

[20] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[21] E. Raymond. The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3):23–49, 1999.

[22] J. Surowiecki, M. Silverman, et al. The wisdom of crowds. *American Journal of Physics*, 75:190, 2007.

[23] S. Wernicke and F. Rasche. FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152, 2006.

[24] G. Wu, M. Harrigan, and P. Cunningham. A Characterization of Wikipedia Content Based on Motifs in the Edit Graph. In *22nd Irish Conference on Artificial Intelligence and Cognitive Science (AICS'11)*, pages 166–173, September 2011.

[25] G. Wu, M. Harrigan, and P. Cunningham. Characterizing wikipedia pages using edit network motif profiles. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 45–52, New York, NY, USA, 2011. ACM.

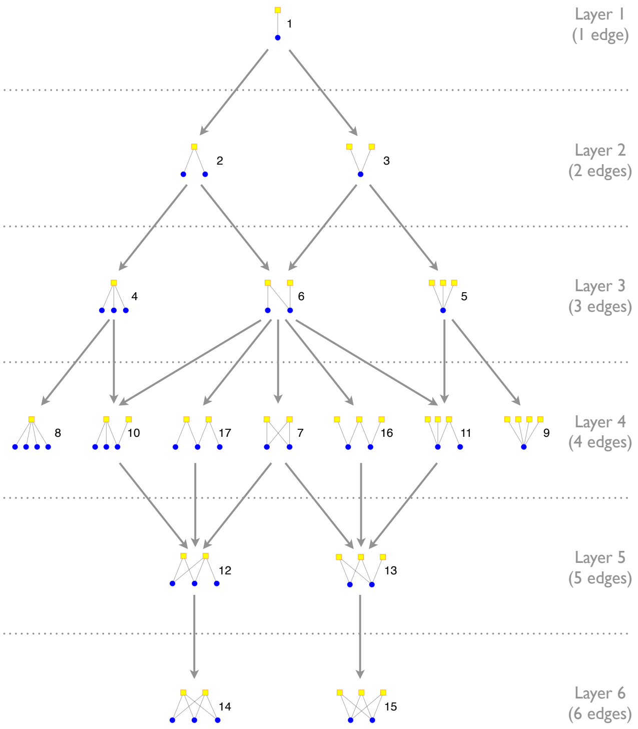

Figure 10: An alternative tree layout for the motifs shown in Figure 2 where the motifs are built up an edge at a time.

APPENDIX

A. ALTERNATIVE MOTIF HIERARCHY

Figure 10 shows an alternative motif hierarchy where the motifs are built up an edge at a time, i.e. each motif contains the nodes and edges of its parent(s) plus the addition of one edge. This tree has a pleasing structure but it has the disadvantage that the counts of parent nodes are not included in the counts of children nodes when only induced motifs are counted as is the case in FANMOD [23] – see Figure 3. For instance, the counts of M6 are not included in the counts of M7, similarly for M11 and M13. An instance of a motif is only included in an induced motif count when *all* of the edges between the nodes occur in the motif.

By contrast, the hierarchy in Figure 2 builds up a node at a time so counts of parent motifs are included in counts of their children.