ASSESSING INFORMATION QUALITY OF A COMMUNITY-BASED ENCYCLOPEDIA

(Completed Paper)

IQ Metrics, Measures, Models, and Methodologies

Besiki Stvilia, Michael B. Twidale, Linda C. Smith, Les Gasser

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, Champaign, IL 61820, USA
{stvilia, twidale, lcsmith, gasser}@uiuc.edu

Abstract

Effective information quality analysis needs powerful yet easy ways to obtain metrics. The English version of Wikipedia provides an extremely interesting yet challenging case for the study of Information Quality dynamics at both macro and micro levels. We propose seven IQ metrics which can be evaluated automatically and test the set on a representative sample of Wikipedia content. The methodology of the metrics construction and the results of tests, along with a number of statistical characterizations of Wikipedia articles, their content construction, process metadata and social context are reported.

1. INTRODUCTION

Information is critical for every aspect of modern life. The quality of information largely determines the quality of decisions made, and, ultimately it affects the quality of activity and action outcomes in organizations and in the society in general. To optimize Information Quality (IQ) assurance activities meaningfully, there is a need to link the changes in IQ to the changes in activity outcomes in a consistent and systematic way. To achieve that however, one has to have a clear and systematic model of assessing IQ first. With the current influx of information the existing culturally justified models of IQ assessment based on careful examination of information by experts and peer groups are breaking down because of cost and scale. There is a need of designing IQ assessment models that are scalable and inexpensive. This paper reports the results of experimentation with a large scale community-based open-access collaborative information collection – Wikipedia[1] - with a goal of designing an IQ assessment model allowing to capture and evaluate a significant portion of its IQ variance automatically.

There are many similarities between Wikipedia and Free Open-Source Software (FOSS) [20] in terms of construction philosophy and methods. FOSS has been around since 1970s with the GNU project and the Free Software Foundation. However, large scale, open access collaborative content creation projects like Wikipedia (the world's largest online encyclopedia) are relatively a new phenomenon. The low barrier of entry, and, consequently the large size and the extreme diversity of its editorial community sharply distinguish Wikipedia from any FOSS project. As of May 17 2005 the English language Wikipedia alone served 1,642,660 pages. In addition to the English language encylcopedia, Wikipedia has encylopedia projects in more than 100 languages[2].

[1] http://en.wikipedia.org/wiki/Main_Page
[2] http://en.wikipedia.org/wiki/Wikipedia:Multilingual_coordination

Wikipedia uses an Internet-based content creation software called wiki[3]. The development of the wiki software is attributed to Ward Cunningham who established the first wiki repository in 1995. The main characteristic of wiki software is extreme reduction in the costs of collaborative content creation, dissemination and upkeep. Wiki makes it possible for anyone with a web browser and access to the Internet to edit the content of an information object and link it to other objects in the collection with very low overhead. At the same time wiki uses a revision control system[4] for logging article edit histories and enabling reverting to earlier versions if needed. This way, the members of the community can monitor and control changes to the article content, again, at very low cost.

With the lack of centralized quality control and the huge size and diversity of the Wikipedia editor population, one may expect an extreme variance in the quality of its articles and lack of trust in Wikipedia as a whole from the public. It has been attested, however, by a number of studies that certain aspects of the Wikipedia IQ such as Currency and the Formality of language are quite high, and Wikipedia has been widely used even by information professionals [12,6]. This paradox calls for closer scrutiny of Wikipedia IQ in general, and its IQ assurance mechanisms in particular. In an earlier paper [15] suggested that the roots of the Wikipedia success might lie in wiki software itself which significantly lowers the cost of collaboration and IQ control, discouraging malicious behavior and low IQ content creation. IQ is a multidimensional construct, however [25]. The quality of Wikipedia articles being high on certain dimensions does not necessarily imply that it is good on the other dimensions as well. Drawing a parallel with FOSS, FOSS has often been praised for stable performance while its usability has not been that good in general [17]. In addition, even though Wikipedia does have a formal set of IQ assessment criteria called the Featured Article Criteria[5], only the IQ of a relatively insignificant number of its articles has been formally evaluated against those criteria. The number of Featured Articles, the articles whose IQ has been confirmed to be meeting the criteria by community votes, was only 236 (<0.05%) out of the total of 545,566 articles in the Wikipedia's collection as of April 21 2005. In addition, the qualitative analysis of the social context of IQ negotiation and control in Wikipedia [22] showed that besides the FA criteria there are a number of additional IQ norms used by the community when evaluating article quality. One such criterion is the norm of an article size not exceeding 32k to ensure a certain level of accessibility over slow dialup modem lines.

One may argue though that as they evolve every Wikipedia article implicitly goes through IQ control based on a peer-review mechanism embedded in the Wikipedia model. The robustness of the Wikipedia peer review mechanism for a given article, however, may depend on the quality of the editorial group of the article and the frequency of edits. It was suggested earlier by [18] that information use improves its quality in general. If this paradigm is true, then the IQ distribution of Wikipedia articles is largely determined by their use and edit distributions. Hence, it is expected the IQ of Wikipedia to be uneven. It would certainly help both Wikipedia editors and common users if the IQ of each Wikipedia article had been explicated through IQ annotations consisting of a set of measurements based on a sound and systematic IQ assessment framework.

This paper explores the ways in which information quality in Wikipedia can be measured in efficient ways. Based on a quantitative analysis of Wikipedia article features and article edit histories we developed a profile of nineteen IQ measures and a set of seven IQ metrics which allow us to assess some of the aspects of the quality of Wikipedia articles inexpensively, that is, automatically. The IQ metrics are tested with two different sets of Wikipedia content for their discriminative power. The methodology of the metrics construction and the results of the tests along with a number of statistical characterizations of Wikipedia articles and the records of their edit histories are reported.

[3] http://en.wikipedia.org/wiki/Wiki
[4] http://en.wikipedia.org/wiki/Version_control_system
[5] http://en.wikipedia.org/wiki/Wikipedia:What_is_a_featured_article

1.1 Overview of Approach

We start with a brief review of the background of the current research and the related past research on Wikipedia and Information Quality. Section 2 introduces the general context of IQ assurance in Wikipedia and the main units of the analysis – articles and their edit history logs. The section also briefly reviews the research design and methodology of this study. Section 3 discusses article profiles of nineteen quantitative measures. It also describes in detail the methodology of IQ metrics construction and the results of their evaluation. We conclude the paper with some wider implications of the work and future research directions.

1.2 Background and Related Research

As an online encyclopedia, Wikipedia draws heavily on the well established genre of printed encyclopedia by importing its form conventions and use norms. Thorough reviews of the encyclopedia genre and its evolutionary history can be found in [4,13]. To the best of our knowledge the most comprehensive framework of encyclopedia quality assessment was proposed by [5]. She defined seven general dimensions of encyclopedia IQ: (1) Scope (Purpose, Subject Coverage, Audience, Arrangement and Style); (2) Format; (3) Uniqueness; (4) Authority; (5) Accuracy (Accuracy and Reliability, Objectivity); (6) Currency; (7) Accessibility (indexing). In addition, two other dimensions – Relevance to user needs and Cost – were defined as contextual, encyclopedia specific. The framework did not include any IQ metrics. This is not surprising, though, as IQ metric functions exploit meta information obtained from an information object/collection and its social context to evaluate the object's IQ indirectly. Therefore, in general, IQ metrics are encyclopedia specific.

Two IQ metrics specific to the Wikipedia context were proposed by [12]: the total number of edits (Rigor) and the total number of unique editors (Diversity). He used the median values of 61 and 36.5 of those metrics as a benchmark when evaluating the quality of Wikipedia articles. [23] reported the values of one IQ measure in their May 2003 analysis of the Wikipedia features. According to them the smallest mean and median article revert times were observed for obscene edits - 1.8 days / 1.7 minutes, and the largest revert times were shown for complete deletions - 22.3 days/90.4 minutes. Interestingly, based on interview data obtained from a small sample of Wikipedia editors, the study reported that the Wikipedia community used reputation-based heuristics to optimize their IQ assurance activities. They examined more carefully the edits made by anonymous users than the edits made by the users with an already established record of "good" edits. [12] found that the most frequent uses of Wikipedia articles in the press were related to current events, slang and colloquial terminology. Extrapolating from these two observations one may suggest that users appreciated the quality of Wikipedia along the Currency and Completeness dimensions, and, at the same time, the expected quality of the edits made by 'heavy' contributors was higher than those made by newcomers. [6] exploited a different kind of article feature to evaluate Wikipedia quality. They counted the frequencies of the parts of speech (POS) known to be characteristic of a formal language genre [3,10] for a sample of 49 Wikipedia articles and compared them to the frequencies of the same POS calculated for the Columbia Encyclopedia. Based on those measurements they concluded that the language of Wikipedia articles is as formal as the language of a printed encyclopedia.

There is a well developed body of IQ research in management science, accounting and the database world. A comprehensive overview of information and data quality assessment tools, frameworks and metrics was given in [24]. [1] proposed the Timeliness metric calculated as the maximum of 0 and 1 minus the ratio of Currency to Volatility where Currency was the age of data plus the delivery time minus the input time, and Volatility was defined as the length of time data remained valid. [14] proposed two IQ metrics for assessing the data quality in relational databases. These dimensions were Soundness and Completeness. Soundness referred to the level of "truthfulness" of the "real world" mapping into a database view, while the Completeness dimension measured the mapping completeness. More specifically, Soundness was the extent to which specific value pairs from the targeted database appeared in the "ideal" database and Completeness, on the other hand, was determined by the extent the pairs from

the "ideal" database appeared in the targeted database. The most relevant set of IQ metrics to the Wikipedia context, however, was proposed by [26]. They defined the following metrics for measuring the quality of webpages: (1) Currency: measured as the function of a page update time stamp; (2) Information-to-Noise Ratio: computed as a ratio of the total length of index items over the page size; (3) Availability: computed as a ratio of the number of broken links to the total number of links in a given web page; (4) Cohesiveness: the degree to which the content of the page is focused on one topic. Cohesiveness was measured with the help of an auxiliary ontology constructed from the Magellan search engine/directory ontology. The Vector Space Model with cosine similarity function and TF-IDF term-weighting [21] was used for assessing the page relevance to the topic vectors in the Ontology. The top 20 matching topics were identified and then based on the level of similarity of these 20 topics, the final Cohesiveness score was calculated; (5) Authority, measured based on the Yahoo's Internet Life Review scores ranging from 2 to 4. If the page had not been reviewed by the Yahoo, then 0 score was given; (6) Popularity which was measured by the number of Web pages citing a particular Web page. The scores were obtained from the Altavista search engine.

Due to the lack of access to information object creation and mediation metadata as well as quality evaluation benchmarks, the earlier IQ assessment frameworks, with the exception of the one proposed by [26], limited themselves to one or two IQ metrics. Fortunately, Wikipedia articles along with the records of their edit histories and the community's quality evaluations embedded in the Featured (best quality) Articles set[6] give us access to this kind of data, the analysis of which will be presented in the next sections.

2. RESEARCH DESIGN AND METHODOLOGY

This section looks at some of the components of the IQ assurance context of Wikipedia which can serve as sources for IQ assessment metadata. The context includes Wikipedia support artifacts, roles and processes. It also briefly reviews the research design and methodology of the study.

2.1 Wikipedia Roles

The Wikipedia context is rich with different roles. Trying to understand those roles and the processes they play in can help us to understand some of the sources of IQ variance in Wikipedia, and consequently the sources of relevant information for IQ metrics. There are at least four distinct roles in the Wikipedia content construction process: (1) Editors: agents that contribute/add new content to the article; (2) Information Quality Assurance (IQA) agents, agents that control the article quality: monitor changes made; revert vandalisms; enhance IQ of the article through minor edits; enhance the IQ of the collection through enforcing IQ norms and criteria across the collection; enhance the IQ of the collection by developing support infrastructure in the form of genre-specific templates and style guides; fostering collaboration and maintaining order in article editorial groups; (3) Malicious agents that purposefully degrade the article quality; (4) Environmental agents that change the representational IQ of articles through changes in the real world states. While mostly degrading IQ, in a few instances Environmental agents may enhance the article's IQ by aligning the real world state with the information contained in an article.

Wikipedia is not a full blown content management system with finely graded user rights/permission management. However, it distinguishes between three groups of accounts: (1) Registered Users – identified and tracked by their login name; (2) Anonymous Users – identified and tracked by the Internet Protocol (IP) address they log on from; (3) Administrators – the same as Registered Users but with special system permissions / privileges. Each of these groups may take different roles at different times. Registered Users can be Editors, IQA or Malicious agents. Likewise, Anonymous Users can act

[6] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

maliciously or can be valuable Editors and IQA agents. Although in theory an Administrator may degrade the article IQ intentionally or unintentionally, one mostly expects "good" edits from them – to play Editor or IQA agent roles.

Wikipedia posts the list of Administrators along with the list of software codes, called Bots, that are run by Wikipedia users for automating some simple IQA tasks such as exporting content from public domain databases, spellchecking or identifying vandalized entries based on a curse word list[7]. As of April 2005 there were 431 Wikipedia users with administrative privileges[8] and the community employed around 50 Bots[9].

Thus, the IQ of a Wikipedia article is constructed through a highly dynamic complex interaction among the members of the above three groups engaged in different activities and playing different roles. The footprint or trace of this interaction is logged and can be obtained from the article's Edit History object which we will discuss next.

2.2 Article Edit Histories

An article edit history is a wiki object that contains the log of three element metadata entries for each instance of the article edit. The edit metadata elements contain the following information: (1) the date and time of the edit; (2) the name of the user who made the edit, or an Internet Protocol (IP) address the edit is made from if the user is not logged on with a Wikipedia registered user name; and, (3) empty or a comment often clarifying the edit purpose. As a result, the history object can be a source of the following meta information about the article: (1) Age, (2) Currency – when the article was updated the last time; (3) The number of times the article has been edited; (4) The names or IPs of the article editors; (5) The types of the edits, such as reverts (returning the article to an earlier state/version), minor edits, copyediting, etc.

The last two kinds of information may not be complete and unambiguous though. It is not necessary for an individual to be logged on or even registered to make an edit. In addition, the same individual can be registered and make edits with more than one name. One can easily extract the set of unique registered user names and IP addresses for anonymous users from the article's edit history. However, this set may not be mapped one to one into the actual set of the article's editors, and can serve only as its approximation. Likewise, edits are often made without editors filling out comment fields, or filling them out with misleading information. As the conventions of commenting the edits are not followed consistently, automatic coding of the comments cannot be fully accurate either. Nonetheless, these two elements of history entries can still provide valuable information about the social structure and dynamics of the article's content creation.

2.3 Featured Articles

Clearly, the main source of Wikipedia article IQ measurements is the article itself. Attributes such as article length, the number of internal and external links, and content readability scores [9] can be extracted and compared across the collection automatically. In addition, the set of Featured Articles[10] (FA) and the FA Criteria[11] can be used for identifying IQ dimensions that the community considers important. They can also be used for obtaining the current IQ requirements and building a baseline and/or target article model for IQ assessment and benchmarking.

FAs are the examples of the Wikipedia's best quality. An article is nominated for FA status by an individual or a group of Wikipedia users, and then the community decides through the process of peer-review and voting whether to feature the article on the Wikipedia main page or not. In a similar way the

[7] http://en.wikipedia.org/wiki/Category:Wikipedia_bots
[8] http://en.wikipedia.org/wiki/Wikipedia:Administrators
[9] http://en.wikipedia.org/wiki/Wikipedia:Bot
[10] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles
[11] http://en.wikipedia.org/wiki/Wikipedia:What_is_a_featured_article

community can strip the article FA status if it no longer meets the FA Criteria. The current version of the FA Criteria consists of eight IQ requirements. An FA has to be: (1) Comprehensive; (2) Accurate and verifiable by including references; (3) Stable - not changing often; (4) Well-written; (5) Uncontroversial – using neutral language and not having an ongoing edit war; (6) Compliance with Wikipedia standards and project guides; (7) Having appropriate images with acceptable copyright status; and (8) Having appropriate length, using summary style and focusing on the main topic.

Interestingly enough, the FA Criteria miss a Reputation dimension. For traditional printed encyclopedias the reputation and expertise of carefully selected editors or groups of editors serve as a guarantee of the article's quality. In contrast to that the Wikipedia IQ assurance mechanism exploits the power of the collective knowledge of a large-scale distributed community following the FOSS quality motto: "given enough eyeballs all bugs are shallow" [20]. Hence, one of the IQ measures can be the number of "eyeballs" – the number of distinct editors. Again this is an indirect measure that happens to be easy to measure. The real number of eyeballs is the number of people reading the article. We are using the number of people bothering to make a change – obviously much smaller and probably more interesting and maybe correlating with the real number of eyeballs.

2.4 Methods

The current study uses statistical analysis and experimentation to identify and model the IQ variable structure of Wikipedia. A sample of 1,000 articles was randomly selected from the 2005/03/09 dump of the English Wikipedia[12] after removing all the redirects to other articles. The qualified population size was 500,623 articles. The sample was further cleaned by removing the articles that contain little content, also known as stubs[13]. As a result the sample size was reduced to 847 articles. One of the original goals of the project was to conduct a longitudinal analysis of the IQ of Wikipedia articles. That meant observing the IQ dynamics of the same sample articles over some period of time. Three consecutive dumps made on 2005/03/09, 2005/04/06 and 2005/04/21 were used for that purpose. To qualify for the analysis, an article had to be present in all those three dumps. This requirement further reduced the size of the sample by 13 articles. Hence, the final size of the Random sample used in this study was 834 articles. In addition, we extracted the titles of the Featured Articles (236 articles as of 04/21/2005)[14] and the histories and discussion pages of both FA and Random sets.

Following the data collection phase, we did statistical analysis of the Random article features and their edit history metadata to construct the profiles of nineteen quality measures. The profiles then were evaluated for redundancy using the technique of Exploratory Factor Analysis [11] and seven IQ metrics were developed. Finally, we evaluated IQ metrics on their ability to capture the IQ variance of the collection by clustering and classifying the labeled profiles of the IQ metrics of the Featured and Random articles. The next section gives a more detailed account of research methodology used and the findings of the analysis.

3. THE RESULTS OF QUANTITATIVE ANALYSIS

To explore the IQ variance structure of Wikipedia content and develop IQ metrics we conducted quantitative analysis of Wikipedia article features and their edit history records. We developed a set of java codes to harvest article edit histories and generate article profiles which combined both the edit history and article feature metadata extracted from the earlier mentioned database dumps. The SPSS implementation of Exploratory Factor Analysis[15] was used to identify groupings of related measures in the profiles. In addition, we used open source software - MySQL RDBMS[16] and WEKA machine learning toolkit[17] to visualize the article profiles and test proposed IQ metrics.

[12] http://download.wikimedia.org/
[13] http://en.wikipedia.org/wiki/Wikipedia:Perfect_stub_article
[14] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles
[15] http://www.spss.com/
[16] http://www.mysql.com/
[17] http://www.cs.waikato.ac.nz/ml/weka/

3.1 Article Profiles

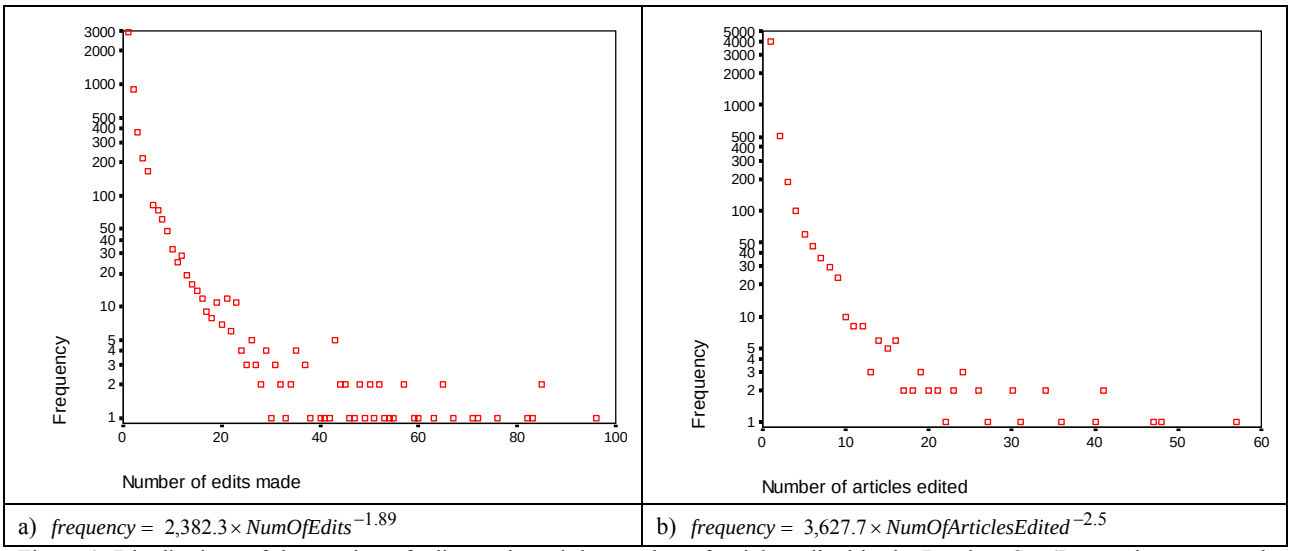
The article profiles contain the descriptive measurements of 19 measures for both the Random and Featured sets. 11 measures out of 19 are based on the article edit history metadata while 8 measures represent the article attributes and surface features (see Table 1).

Embodying the community IQ valuations, FAs along with the FA criteria are very valuable resources for designing and validating IQ metrics as well as constructing a benchmark for article IQ assessments. A brief look at the profiles makes clear that the median values of the measures vary widely for the Featured and Random sets. The median length of the Featured articles is 18 times larger than that of the Random set. The articles also vary sharply on the Total Number of Edits and the Number of Images. 652 (78%) articles from the Random set do not contain any image, while only 5 (2%) articles of the Featured set are without images. The median numbers are also significantly different: 5 images for the Featured Set vs. 0 for the Random Set. Furthermore, the medians of the Number of Edits and Age variables for the Featured articles are much higher than for the Random sample. Likewise, the Featured articles show relatively better document surface readability scores but higher redundancy scores: Flesch - 36 vs 27 of the Random sample (higher is better); Kincaid - 12 vs 13 of the Random sample (lower is better); and Informativeness / Redundancy – 0.52 vs 0.32 of the Random sample. Redundancy here is calculated using the Information Noise metric originally proposed by [26].

Not surprisingly, the Featured articles are reverted back faster than the articles from the Random set showing the mean/median revert time of 199/9 minutes for the Featured articles vs. 712/20 minutes of the Random articles. These statistics were extracted from the aggregate edit histories of each set where the edits made on different articles in the set were not distinguished from one another. Consequently, one may expect the statistics to be biased towards the articles with a high number of reverts. Indeed, the median value of the Article Median Revert Time measures of the Random Set article IQ profiles is 0 as most of the articles in that set had never been reverted (see Table 1)

Even though the mean and median times of 'dirty states' – the state between a revert and the edit immediately before it – is relatively small (712/20 minutes) for Wikipedia in general, it still points to a significant reliability problem for the encyclopedia as the users accessing an article during those 712/20 minutes may not know that the article is in an invalid state. Furthermore, while a high update rate is beneficial for keeping article content up-to-date, especially those related with current events, the stability and verifiability of articles may suffer. There is no guarantee that the content the user makes a reference to will be found next time.

The statistics also suggest that the Number of Edits may have a power law distribution. 6% of the editor pool (Administrators) do 24% of the edits for the Random set articles. This ratio is even higher for the Featured articles – 2% over 21%. The relationship between the Number of Edits and the Frequency of Editors with that number of edits, can be modeled with the following power law formula: ***Number of Editors with n Edits* = b*(n^(-k)).** For the Random Set the SPSS implementation of least square regression with a power law model estimates the values of the above coefficients as b= 2,382.3 and k = 1.89 (R^2 = 0.93, F= 859, sign. 0.0000). Likewise, the Number of Distinct Articles Edited and the Frequency of Editors with that number of distinct articles edited shows a power law distribution as well - ***Number of Editors with n Number of Articles Edited* = 3,627.7*(n^(-2.5))** (R^2 = 0.98, F= 916, sign. 0.0000) (see Figure 1,2)**.** This echoes the findings of the study of Open Source Software communities by [16] which showed that the author productivity patterns in Linux Software Map (LSM) and Sourceforge followed a power law – few contribute to many and many contribute few. In particular they estimated the values of **k** for LSM and Sourceforge as 2.82 and 2.55.



a) $frequency = 2{,}382.3 \times NumOfEdits^{-1.89}$

b) $frequency = 3{,}627.7 \times NumOfArticlesEdited^{-2.5}$

Figure 1: Distributions of the number of edits made and the number of articles edited in the Random Set (Log scales were used for Y axis).

Hence, if we consider Wikipedia editors as vertices in a graph and articles as edges connecting the editors who edited the same article at least once, we may have a scale-free network [2,8]. A scale-free network is a network where the distribution of connectivity is uneven and follows a power law. The plot (see Figure 1,b) shows that 4,018 editors edited only one article while only 5 editors edited more than 40 articles. The total number of articles in the Random sample was 834 and the total number of distinct/unique editors is 5,100.

The main characteristics of scale-free networks are growth and preferential attachment. The English Wikipedia does exhibit both characteristics. Its growth rate has been remarkable. The number of articles that were not redirects in the 03/09/05 dump was 500,623. This number grew to 528,291 for the 04/06/05 dump and to 545,566 for the 04/21/05 dump. Consequently, the growth of the English Wikipedia article collection between 03/09/05 and 04/21/05 was 545,566-500,623 = 44,943 ≈ 9%. We did not have access to the complete list of Wikipedia editors. However, the analysis of the combined log of the edit histories of the Random set articles shows that the number of unique editors increased from 4,571 to 5,100 between 03/09/05 and 04/21/05. Hence, the growth of the number of the unique editors in 43 days was 529 (≈12%). At the same time, Wikipedia has a few important members carrying out a large load in terms of content creation and quality maintenance, and, many members or anonymous editors making only one or two edits. Interestingly enough, out of the 25 top most "productive" (by the number of articles edited) editors in the Random Set only 15 were administrators.

3.2 Metrics

As stated above the main goal of extracting article profiles was to develop a set of IQ metric functions that could be used for evaluating IQ dimensions of the Wikipedia articles. The IQ metrics should allow generating measurements that can capture a substantial portion of IQ variance of an article, and at the same time avoid redundancy and bias. This implies the IQ measurements to be correlated as little as possible. However, the edit history based measures in the profiles are highly correlated. The same applies to readability and revert related measures. To lose the least amount possible IQ variance imbedded in the article profiles, we decided not to discard the correlated measures but group them together in less correlated IQ metrics functions.

We applied a technique of exploratory factor analysis to 834 article profiles from the Random set to identify variable groupings [11]. The factor analysis suggested seven IQ metrics based on the variable groupings identified by the first seven extracted components. We decided to retain extracted component score coefficients when defining the IQ metrics functions since the profile measures used different scales and were not normalized:

1. ***Authority/Reputation*** *= 0.2 * Num. Unique. Editors + 0.2 * Total Num. Edits + 0.1 * Connectivity + 0.3 * Num. of Reverts + 0.2 * Num. External Links + 0.1 * Num. Registered User Edits + 0.2 * Num. Anonymous User Edits.*
2. ***Completeness*** *= 0.4 * Num. Internal Broken Links + 0.4 * Num. Internal Links + 0.2 * Article Length*
3. ***Complexity*** *= 0.5 * Flesch Readability Score – 0.5 * Kincaid Readability Score*
4. ***Informativeness*** *= 0.6 * InfoNoise - 0.6 * Diversity + 0.3 * Num. Images*
5. ***Consistency*** *= 0.6 * Admin. Edit Share + 0.5 * Age*
6. ***Currency*** *= Currency*
7. ***Volatility*** *= Median Revert Time*

We tested the power of the IQ metrics for capturing the significant portion of IQ variance in the collection by computing seven IQ measurements for each article in the pooled set of the Featured (236 articles) and Random (834 articles) sets, and then clustering and classifying the articles along those measurements.

As the majority of the original attributes as well as the derived IQ measurements were not normally distributed and exhibited non-linear patterns, we decided to use the WEKA implementation of the Density Based Clustering algorithm. The Density Based Clustering algorithm is known to perform well in identifying non-linear, spatially distributed clusters [7]. Indeed, it did better in comparison to the K-Means Clustering algorithm when clustering the Wikipedia sets based on the measurements. The classes to clusters comparison showed that only 152 (14%) articles were put in incorrect clusters. It is important also to note that most of these articles were from the Random set. Only 22 articles from the Featured set (9%) were not clustered correctly. For K-Means these numbers were 268 (25%) and 46 (19%) respectively.

After clustering we used supervised classification to further test the IQ metrics. In particular, we applied the WEKA implementation of the C4.5 Decision Tree classification algorithm [19] to the pooled collection of the Featured and Random articles labeled by the set names (Featured and Random). With 10 fold cross-validation the precision and recall for the Featured set were 90% and 92% respectively. For the Random set these numbers went up to 98% and 97%. Interestingly enough, the decision tree misclassified the profile of the H II Region[18] from the Random Set as Featured - an article which belonged to both sets and was the only overlap between them. It is clear from the median values of the misclassified subset of the Random Set that those articles are much closer to the center of the Featured Set than the center of the Random Set (see Table 1). This suggests a potential use of misclassified profiles for identifying articles that have already reached a certain level of maturity in their IQ and can be nominated for FA status.

Thus, our IQ metrics have been shown to be successful in discriminating high quality articles in the Wikipedia collection based on article features and edit history metadata. As the Wikipedia collection changes over time along with the composition of the FA set and FA criteria, we may need to rerun the experiment to see if the above performance figures will still hold. The fact that they can be calculated automatically is a clear advantage of these metrics. On the con side, however, the metrics are not "deep" and can not assess the semantic quality dimensions of article content. The median values of the Featured Set IQ metrics can be used as target values in assessing the IQ of the rest of the Wikipedia article collection (see Table 2). To optimize IQ assurance activities, however, we may need to identify critical values for each IQ measure (see Table 1) and the relationships of those critical values with related IQ

[18] http://en.wikipedia.org/wiki/H_II_region

metrics they participate in and the overall IQ valuations. These critical values can be obtained through the qualitative analysis of IQ negotiation and peer-review discussion instances accompanied with the quantitative analysis of the features of "low" quality articles identified as such by the community.

4. CONCLUSION

In this paper we presented a methodology of IQ metrics construction and validation for Wikipedia. A set of seven IQ metrics was developed and tested on a representative sample of the collection. A number of statistical characterizations of a random sample of articles were reported and interpreted.

The analysis showed that the open access to Wikipedia articles, their content construction and IQ assurance process metadata allows us to capture and evaluate a significant portion of Wikipedia IQ variance in an inexpensive and scalable way. The proposed metrics were successful in indirectly assessing Wikipedia article quality along the dimensions corresponding to the set of IQ criteria (requirements) adopted by the community. The analysis of the IQ metrics scores showed that they were clearly able to discriminate the high quality articles voted for by the community from the rest of the collection. In addition, the methodology used for IQ metrics construction allows us to suggest that the IQ metrics will be robust to possible future changes in the community IQ assessment criteria, and the methodology can be reused and/or adapted for IQ evaluation and maintenance in many other similar contexts of information content creation.

In future research we plan to continue analyzing the Wikipedia IQ assurance context both qualitatively and quantitatively to gain a better understanding of how IQ choices and assessments are made by the Wikipedia community; how the community's social network topology may affect those choices; and how they can be approximated by quantitative measures and captured into baseline and target models of IQ assessment.

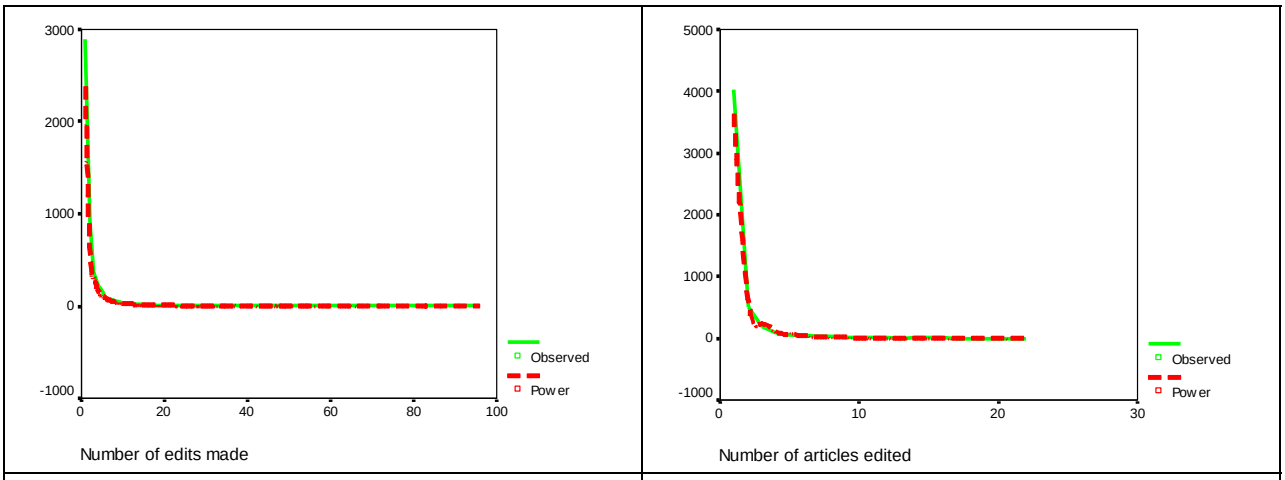
APPENDIX:

Measures	Featured Set (236 articles)	Random Set (834 articles)	The Set of Misclassified instances of the Random Set (10 articles)	Source
Num. of Anonymous User Edits	82	2	87	Edit History
Total Num. of Edits	257	8	251	Edit History
Num. of Registered User Edits	171	6	149	Edit History
Num. of Unique Editors	108	5	109	Edit History
Article length (in # of characters)	24,708	1,344	20,949	Article
Currency (the time between the dump date and the date of the last update of the article) (in days)	3	46	2	Edit History
Num. of Internal Links	206	17	176	Article
Num. of Reverts	12	0	8	Edit History
Num. of External Links	9	0	12	Article
Article Median Revert Time (in Minutes)	9	0	17	Edit History
Num. of Internal Broken Links	6	0	8	Article
Article Connectivity (# of Articles connected to a particular article through common editors)	836	154	826	Edit History
Num. of Images	5	0	2	Article
Article Age (in days)	1,153	388	1,153	Edit History
Diversity (# of Unique Editors / Total # of Edits)	0.4	0.7	0.4	Edit History
Information Noise*(content) = 1- (The size of the term/token vector after stemming and stopping)/(document size before processing)*	0.52	0.32	0.54	Article
Flesch	36	27	36	Article
Kincaid	12	13	12	Article
Admin. Edit Share (Num. of Admin Edits / Total Num. of Edits)	0	0	0.037	Edit History

Table 1: Descriptive statistics of article profiles (medians of the article IQ profile values).

IQ Metrics	Featured	Random
Authority/Reputation	198.1	19.8
Completeness	5,014.2	275.6
Complexity	11.8	6.9
Informativeness	1.4	-0.2
Consistency	576.5	194.0
Currency	3	46
Volatility	9	0

Table 2: Median values of the IQ Metrics.



a) $frequency = 2{,}382.3 \times NumOfEdits^{-1.89}$
$R^2 = 0.93$, F= 859, sign. 0.0000

b) $frequency = 3{,}627.7 \times NumOfArticlesEdited^{-2.5}$
$R^2 = 0.98$, F= 916, sign. 0.0000

Figure 2: Distributions of the number of edits made and the number of articles edited in the Random Set (power curves are fitted using the SPPS least square regression curve fitting utility).

REFERENCES:

[1] Ballou, D., Wang, R., Pazer, H., Tayi, G. (1998). Modeling information manufacturing systems to determine information product quality. Management Science, 44(4), 462-484.

[2] Barabasi, A., Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.

[3] Biber, D. (1988). Variations across speech and writing. Cambridge, UK: Cambridge University Press.

[4] Collison, R. (1966). Encyclopaedias: their history throughout the ages (2 ed.). New York, NY: Harper.

[5] Crawford, H. (2001). Encyclopedias. In: R. Bopp, L. C. Smith (Eds.), Reference and information services: an introduction (3 ed.). (pp. 433-459). Englewood, CO: Libraries Unlimited.

[6] Emigh, W., Herring, S. (2005). Collaborative authoring on the Web: a genre analysis of online encyclopedias. In: Proceedings of the 39th Hawaii International Conference on System Sciences.

[7] Ester, M., Kriegel, H. P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases. In: Proc. of the Second International Conference on Data Mining KDD-96. Portland, OR, 226-231.

[8] Goh, K., Oh, E., Jeong, H., Kahng, B., Kim, D. (2002). Classification of scale free networks. PNAS, 99(20), 12583-12588.

[9] Gunning, R. (1952). Technique of clear writing. .McGraw-Hill.

[10] Heylighen, F., Dewaele, J. (2002). Variation in the contextuality of language: an empirical measure. Foundations of Science, 6, 293-340.

[11] Johnson, R., Wichern, D. (1998). Applied multivariate statistical analysis (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

[12] Lih, A. (2004). Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In: Proceedings of 5th International Symposium on Online Journalism. Austin, TX.

[13] McArthur, T. (1986). Worlds of reference: lexicography, learning and language from the clay tablet to the computer. Cambridge, UK: Cambridge University Press.

[14] Motro, A., Rakov, I. (1998). Estimating the quality of databases. In: Proceedings of FQAS 98: Third International Conference on Flexible Query Answering Systems. Roskilde, Denmark.

[15] Neus, A. (2001). Managing information quality in virtual communities of practice: Lessons learned from a decade's experience with exploding Internet communication. In: Proceedings of the 6th International Conference on Information Quality. Boston, MA.

[16] Newby, G., Greenberg, J., Jones, P. (2003). Open source software development and Lotka's Law: Bibliometric patterns in programming. Journal of the American Society for Information Science and Technology, 54(2), 169-178.

[17] Nichols, D., Twidale, M. (2002). The Usability of Open Source Software. First Monday, 8(1).

[18] Orr, K. (1998). Data quality and systems theory. Communications of the ACM, 41(2), 66-71.

[19] Quinlan, J. (1993). C4.5: Programs for machine learning. San Francisco, CA: Morgan Kaufmann.

[20] Raymond, E. (1998). The cathedral and the bazaar. First Monday, 3(3).

[21] Salton, G., McGill, M. (1982). Introduction to modern information retrieval. New York: McGraw-Hill.

[22] Stvilia, B., Twidale, M. B., Gasser, L., Smith, L. C. (2005). Information quality in a community-based encyclopedia. Submitted to the International Conference on Knowledge Management - ICKM 2005.

[23] Viegas, F., Wattenberg, M., Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In: Proceedings of CHI 2004. Vienna, Austria, 575-582.

[24] Wang, R., Allen, T., Harris, W., Madnick, S. (2003). An information product approach for total information awareness. In: Proceedings of IEEE Aerospace Conference.

[25] Wang, R., Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-35.

[26] Zhu, X., Gauch, S. (2000). Incorporating quality metrics in centralized distributed information retrieval on the World Wide Web. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 288-295.