Journal of Web Engineering, Vol.3, No.3&4 (2004) 216-235

RANKING SEARCH RESULTS BY WEB QUALITY DIMENSIONS

JOSHUA C. C. PUN

Hong Kong University of Science and Technology
punjcc@cs.ust.hk

FREDERICK H. LOCHOVSKY

Hong Kong University of Science and Technology
fred@cs.ust.hk

Received July 15, 2004
Revised November 15, 2004

Currently, search engines rank search results using mainly link-based metrics. While usually most of the search results are relevant to a user's query, due to how the results are ranked, users often are still not totally satisfied with them. Using a proposed framework of web data quality, it is found that current search engines usually only consider a very small number of the dimensions of web data quality in their ranking algorithms. In this paper, a newly identified web data-quality dimension, *appropriateness*, which is based on the linguistic and visual complexity of a web page, is studied. It is computed using new metrics that classify web pages into three main appropriateness genres: scholarly, news/general interest and popular. Experiments have shown the effectiveness of the metrics in ranking web pages by whether they are appropriate to a user's task and information needs.

Key words: Web data quality, web metrics, user appropriateness
Communicated by: A Spink and C Watters

1 Introduction

Currently, primarily link-based metrics are used by most search engines to rank the search results. While most of the results returned are usually relevant to the query, this may not be sufficient to satisfy a user's task and information needs, since the pages most appropriate to the user's needs may not be ranked near the top of the results. For example, if a user submits the query "SARS" to a search engine, his intent may be to find basic information on what SARS is and how to guard against being infected. However, the search engine may return, as the highest ranked results, technical pages that discuss in-depth the virology, transmission, epidemiology, prevention and treatment of SARS in medical terms. While the pages in which the user is interested may be present somewhere in the results returned, they may not be immediately and easily accessible to the user. This example illustrates that, in general, the issue of ranking results in a way that satisfy the user's task and information needs is not addressed well by current ranking strategies of search engines and, thus, the quality of the search results may not meet the user's expectations (e.g., whether the results are useful and comprehensible to the user).

In this paper, we examine a missing link between the relevance and the quality of the search results returned to the users. Relevance does not necessarily imply good quality. It is possible to have relevant, but poor quality, search results. Current search engines (e.g., Google) only take one of the web data-quality dimensions into consideration (i.e., the *believability* dimension based on the use of

hyperlink-based metrics for computing the reputation and authority of web pages). They seldom employ the remaining dimensions.

We propose a framework of web data-quality dimensions that incorporates the essential results from previous studies on data quality and examines the applicability of each data quality dimension to web data. In addition, the framework also incorporates some new dimensions, which are either not considered at all or considered not to be so important for traditional data. For example, one possible unmatched user expectation for "good quality" search results arises from the issue of whether the results returned by a search engine are *appropriate* for the user's task and information need. For traditional data, appropriateness is seldom considered since usually the authorship of and access to the data is tightly controlled. People for whom the data is not appropriate normally would not have easy access to it (or even bother to access it). On the Web, however, editorial control is often lacking and access normally is available to anyone with a browser and Internet connection. In view of this, the issue of appropriateness, particularly in terms of search results, becomes relevant and, hence, a new data quality dimension "appropriateness" is added to the framework. In this paper, we examine how this new data quality dimension can be used to improve the ranking of web search results.

The appropriateness of a web page can be defined in many ways and can depend on many factors. For example, it may depend on the expected level of detail or on the use of words in the page. For a given topic, a layman and a specialist often have quite different requirements for their task and information needs. While a layman may be very satisfied with a general article on a newspaper web page, a specialist might prefer to read a web page from a professional organization or a paper from a prestigious journal. Nonetheless, the specialist still may be satisfied with a general article. It all depends on why he asks a particular query. Hence, whether a web page is an appropriate response to a user query needs to be considered in light of the task for which the information is required.

Consequently, to determine the appropriateness of a web page, the intended audience and use of the web page needs to be known. In this paper, to determine the intended audience and use of a web page, we analyse its linguistic and visual complexity to classify it as one of three main (appropriateness) genres: *scholarly, news/general interest* or *popular*. For each genre, some language- and visual-based attributes of web pages are used to characterize it. Our approach is to measure these attributes automatically and use them to estimate the likelihood that a web page is for a particular intended audience type and use (i.e., is one of the three appropriateness genres). Then, when users submit their queries to a search engine with their preferred appropriateness genre, the search engine can provide them with both relevant *and* appropriate pages.

The rest of this paper is organized as follows. The next section briefly describes related work on data quality, evaluation of medical information on the Web and web genre. Section 3 presents a framework of web data quality and some new dimensions of web data quality. Section 4 explores one of these new dimensions, *appropriateness*, in further detail by discussing a methodology for measuring it in terms of three new metrics and the issues involved in their implementation. In Section 5 we present our experimental results on the effectiveness of the metrics in classifying web pages according to their appropriateness genre. Finally, Section 6 concludes the paper and identifies some possible future research directions.

2 Related Work

2.1 Dimensions of Data Quality

To devise a framework for web data quality, previous studies in the information systems and database fields on the dimensionality of data quality can be examined, as web data is also a kind of data. Fox *et al.* [9] lay a foundation for the study of data quality. They discuss five different approaches to define data and propose data quality dimensions where the most important dimensions are: *accuracy, completeness, consistency* and *currentness*. Yang *et al.* [28] continued the Fox *et al.* study and classified data quality with different dimensions: *accessibility, interpretability, usefulness* and *believability*. They further classify each of these dimensions into a number of facets. For example, accessibility has the facets *availability* and *access security*. Interpretability has the facets *syntax* and *semantics*. Usefulness has the facets *relevance* and *timeliness* (*current* and *non-volatile*). Lastly, believability has the facets *completeness, consistency, credibility, accuracy* and *objectivity*.

Strong *et al.* [20] define high-quality data as data that is fit for use by data consumers. This means that *usefulness* and *usability* are important aspects of quality. Using this definition, the characteristics of high-quality data consist of four data quality (DQ) categories: *intrinsic, accessibility, contextual* and *representational*. Each category has different data quality dimensions. For example, *Intrinsic DQ* has the dimensions accuracy, objectivity, believability and reputation. *Accessibility DQ* has accessibility and access security dimensions. *Contextual DQ* has dimensions relevancy, value-added, timeliness, completeness and amount of data. Finally, *Representational DQ* has dimensions interpretability, ease of understanding, concise representation and consistent representation. All these dimensions give a broader conceptualization of data quality than the conventional view, which only focuses on intrinsic aspects of data quality and fails to address the broader data quality concerns of data consumers.

2.2 Evaluation of Medical Information on the Web

Poor quality medical information found on websites has, perhaps, the most serious impact of all kinds of data, as it deals with human health and life. This issue has aroused concern from the health and medical communities. The Web gives patients and health care professionals access to millions of pages of clinical information but, at the same time, it is getting more difficult for web users to judge the reliability and quality of health and medical information found on the Web. Suggestions and guidelines [5, 26, 27] have been proposed from medical associations and professionals on how to prepare and evaluate a quality medical website. Ultimately, the quality of the content still depends on the same factors that readers of print publications rely on: authorship of the content, attribution to the sources of content, disclosure of funding and competing interests, and timeliness of the information presented. In addition, the "scope and intended audience" of medical information on a website (i.e., whether it is for a layperson or a professional) also is an issue that needs to be considered when making use of the information.

2.3 Classification of Web Pages / Web Genre

A genre is a “classifying statement” that allows us to recognize items that are similar even in the midst of great diversity [18]. Various ways of classifying the genre of web pages have been proposed based primarily on the purpose of the web pages. Cornell University Library [7] has four broad classification categories: *scholarly, substantive news/general interest, popular* and *sensational*. On the other hand, Alexander and Tate [1] have proposed six classification categories: *advocacy, information, news, personal, business* and *entertainment*. The major distinction between these two classifications is that

the former one is more focused on the *readability* of a web page whereas the latter one is more focused on the *functionality* of a web page. Apart from these two commonly used classifications, there is also a classification used by the Webby Awards that is also focused on functionality, but it is even more specific and has 30 categories [25].

To improve communication and sharing of resources, the ability to identify the genre of a web page is very important. Crowston *et al.* [8] examined 100 web pages with the intention to look for reproduced and emergent genres. On the basis of form and purpose, they identified 48 different genres. In 2001, Roussinov *et al.* [17] did a larger study of genre on the Web with 1234 web pages. There were 116 different genres identified. However, one major outstanding issue in this work is that the genre of a web page could not be recognized automatically.

There are also various traditional ways of classifying web pages (e.g., Yahoo and Open directory project). Generally, these classifications of web pages are *data-oriented* as they are based on the subject (or main theme) of the data or web pages. Our classification of the appropriateness of a web page, which is based on the web page's linguistic and visual characteristics, is different in that it is *user-oriented* taking into consideration the user's task and information needs. Moreover, this classification is orthogonal to traditional classifications since, given a set of web pages under the same category (genre) in a traditional classification, they can be further classified by our classification.

3. Web Data Quality

3.1 Framework of Web Data Quality

With reference to the guidelines recommended by library science [21] and information systems [29] researchers, as well as an understanding of the intrinsic difference between traditional data and web data, we have identified six additional data quality dimensions, namely, *navigation, visual appearance appropriateness*, *cohesiveness, minimality* and *popularity* that are particularly applicable to web data. Together with the dimensionalities of data quality from previous work, these form a framework for web data quality as shown in Figure 1. In the rest of this section, we elaborate on the applicability of existing data quality dimensions to web data as well as discuss the newly identified data quality dimensions.

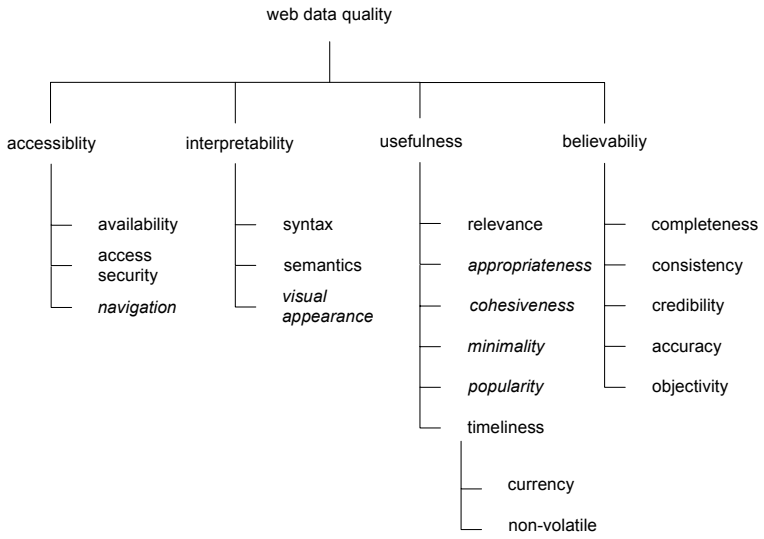

Figure 1. A framework of web data quality.

3.2 Applicability of Data Quality Dimensions

Not all dimensions of data quality are necessarily directly applicable to or meaningful for web data. Furthermore, the meaning of some of these dimensions may need to be redefined for web data. For web data, the *availability* dimension can be interpreted as whether web pages can be accessed free of charge (i.e., publicly available). However, some pages are restricted to registered users only (e.g., Web pages that require a login by their users). Web *accessibility*, on the other hand, can refer to the universality of free access of web pages by everyone regardless of their (dis)ability. *Interpretability* relates to the comprehensibility of web pages. By using more tables, point form, headings and properly emphasized text and colour, a web page may improve its interpretability. *Usefulness* of web data relates to the traditional information retrieval area where a determination is made of how *relevant* search results are to the user query. For web data, *believability* (and *credibility*) relate to the *reputation* and *authority* of web pages. They are commonly used in search engines following the introduction of the link-based metrics PageRank [4] and Hubs/Authorities [13]. These metrics have also been interpreted as a measure of "page popularity" [6] and "page quality" [2, 3]. Table 3 in the Appendix summarizes the various data quality dimensions previously proposed.

3.3 Navigation

The navigation dimension is related to the features available for moving around a web page/site and, thus, is one measure of a page's/site's accessibility. A good web page/site should help the user find the information he wants and let the user feel comfortable to navigate the page/site. Therefore, this is a measure of the extent to which the web page/site is easy to navigate and to apply to different tasks. Ease of navigation, can be achieved by providing indications of the user's location within a web site, navigation aids, directions for navigating a web site, etc.

3.4 Visual Appearance

The visual appearance dimension is related to features having to do with the aesthetics of a web page and, thus, is one measure of a page's interpretability. A web page with a good visual appearance is

more likely to draw and keep the users' attention and help him understand the information being presented. Visual appearance relates to features as simple as consistency in the page layout or to more complex features (e.g., attractiveness of the screen layout, background and pattern, overall use of colour, sharp displays, adequate brightness of pages, presence of eye-catching images or title on the homepage, etc.). This dimension is more subjective and more complicated to measure than the accessibility dimension, which is based primarily on the recommendations of the Web Accessibility Initiative (WAI) on the use of colour [23].

3.5 Appropriateness

Appropriateness is a measure of how well the content of a page matches the user's task and information needs[a] and, thus, is one measure of a page's usefulness. Given a user need, the appropriateness of a web page can depend on many factors. For example, it may depend on the expected level of detail or on the use of words in the page. For a given topic, either a general article on a newspaper web page or a paper from a prestigious journal could be appropriate depending on the user's task and information needs. Therefore, to find highly suitable content, a good understanding of the user as well as the user's current task is needed. For example, which reading level is appropriate for the user and task? Which content is appropriate for the user's age, knowledge level and task?

The rationale for this dimension is that when authors write something, in whatever format (e.g., books, articles or web pages), they normally have an intended audience and use for the data in mind. The intended audience and use reflects the purpose of their writing, for example, information dissemination, education and training, commerce and advertising, or entertainment and communications to name a few. If the type of reader and his task and information needs are a good match with the intended audience and use, then the page is appropriate for that reader and he will enjoy reading the page; else, the mismatch may result in a disappointing and unhappy reading experience.

3.6 Cohesiveness

The cohesiveness dimension gauges how closely the concepts are related to each other in a web page and, thus, is one measure of a page's usefulness. In software engineering, a good programming practice is to write a cohesive software module. Using the same rationale when writing a document or a web page, a good writing practice is to write a cohesive page. If the concepts in a web page are highly diversified and unrelated, the cohesiveness of the page is weak. If not, the cohesiveness is considered to be strong. To be of good quality, a web page should be cohesive. If two web pages are equally relevant to a query, a user would probably prefer to read a more cohesive web page since it would be more focused on the topic and thus more useful to him. The cohesiveness dimension is a self-contained concept and does not relate to other factors (e.g., the query). As such, it is possible to have a highly cohesive page returned that is totally unrelated to the query! Therefore, what we want is a page that is highly cohesive as well as relevant to the query.

---

[a] Appropriateness can be viewed from two perspectives: *queries* and *users*. For queries, appropriateness (usually called relevance) considers how well the retrieved web pages match with the query conditions. For users, on the other hand, appropriateness as defined here, considers how well the retrieved web pages match with the user's linguistic and visual needs given a certain task and information need.

3.7 Minimality

The minimality dimension is defined as the proportion of useful information that is contained in a web page and, thus, is one measure of a page's usefulness. A web page can contain a lot of text and images. Often, not all of this information is useful to users. For example, navigation hyperlinks that appear on every page are usually not useful information from an information content perspective[b]. As much as possible, everything on a web page should be useful (i.e., we want to minimize the reading of unnecessary information). The measurement of this dimension for a page will, therefore, give an indication of how much "real" information content there is in the page. If there are two web pages that have similar real information content, then one would probably prefer to read the web page that is shorter, since it will take less time to digest the material.

3.8 Popularity

The popularity dimension considers how popular a web page is to all users on the Web and, thus, is one measure of a page's usefulness. The more people that read a page, the more "useful" one would think that this page is most of the time. Popularity can be interpreted as the number of hits or visits per page. This figure, however, is only available to the page owner itself. While it could be calculated by examining the log file of a web site, this is an unreliable indicator of page popularity. In [10], a refined metric for popularity is proposed which takes into account structural information. Another possible way to measure this dimension is to measure it indirectly by counting how many other web pages have cited this page.

4. Determining the Appropriateness of Web Pages

In this section, the appropriateness dimension will be further explored. For our purpose in determining the appropriateness of a web page, the emphasis is on identifying a broad category of the intended audience and use of a web page based on its readability (i.e., linguistic and visual complexity). Its precise functionality will generally not be considered. The main focus of the classification proposed by Cornell University Library [7] is therefore adopted. However, this classification is generalized into only three genres: *scholarly, news/general interest* and *popular* where the sensational genre is dropped and generalized to the popular genre as, generally, for both genres their purpose is to arouse interest and sell ideas or products.

4.1 Overview of Classification Methodology

To classify the appropriateness genre of web pages as one of scholarly, news/general interest or popular, we use a *support vector machine* (SVM), which is a machine learning technique that integrates dimension reduction and classification [22]. A classification task usually involves training and testing data, which consist of some data instances. Each instance in the training set contains one *target value* (class label) and several *attributes* (features). The goal of a SVM is to produce a model that predicts the target value of data instances in the testing set given only the attributes. Unlike other machine learning methods, the performance of a SVM is not related to the input dimensionality, but to the *margin* with which it separates the data. Since, experimentally, SVMs have achieved superior

---

[b] Hyperlinks may improve the navigability of a web page and some of them may point to relevant content. However, from an information content perspective, these links are considered to be not (or less) important [19].

performance on a number of high-dimensional data sets (e.g., automated text categorization [14]), we employ them here.

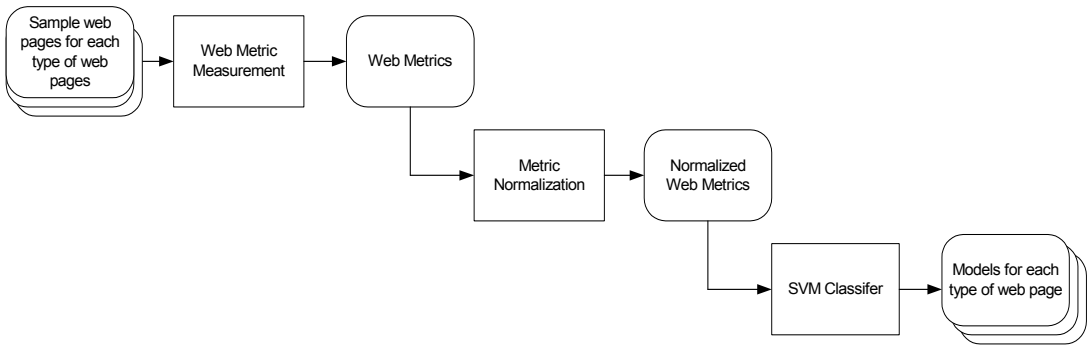

Figure 2. SVM Model Training

The general methodology that we use to classify the appropriateness genre of web pages using a SVM is as follows:

1. A sample data set is built and analyzed to understand the characteristics (or distribution) of the attributes for a given appropriateness genre. Based on their characteristics, the values of the attributes are normalized.
2. The normalized values of the attributes serve as inputs to a SVM for model training (see Figure 2) on each appropriateness genre.
3. The models trained (i.e., the new metrics developed) for each appropriateness genre are used to predict the likelihood that the testing data (or web page) is of that genre.
4. The higher the value returned from the model for a particular genre for the testing data, the more likely the testing data will be classified as that genre.

4.2 Characteristics of Attributes

To understand the characteristics of the attributes for classifying a web page according to its appropriateness genre, a sample data set was analysed. From this data set, various aspects of the attributes were measured to obtain the following values:

- A_i denotes attribute *i* for a web page (e.g., Fog index is one of the attributes).
- $\overline{A_i}$ denotes the mean of attribute *i* for all web pages
- A_i^j denotes that for attribute *i* of a given appropriateness genre, the top $j \times 100\%$ of web pages (ranked in descending order of attribute *i*) have a value higher than A_i^j, where *j* is *0.2, 0.5* and *0.8*

The values of $A_i^{0.2}$, $A_i^{0.5}$ and $A_i^{0.8}$ are determined as follows. For each attribute *i*, web pages in the data set are ranked in descending order and the distribution of their values is obtained (shown in Figure 5, Figure 6 and Figure 7 in the Appendix). While the actual top three *x* percentages for the three turning-point (or cut-off) values for each attribute may not be the same, determining the cut-off percentages for each cut-off value of the different attributes is too complicated and is over-specified. Furthermore, instead of selecting the three cut-off values evenly, say 25%, 50% and 75%, we instead

select cut-off values that are more towards both ends of the distribution to better reflect the distribution of the values for each attribute. Consequently, all attributes use the same cut-off percentages at the top 20%, top 50% and top 80%. Hence, the values of $A_i^{0.2}$, $A_i^{0.5}$ and $A_i^{0.8}$ are determined accordingly.

4.3 Relationship and Normalization of Attributes

Before normalization, the relationship between an attribute and a particular appropriateness genre (i.e., whether it is directly, inversely or not correlated) is identified by correlation analysis. If the higher the value of an attribute the more likely the web page is of a particular genre, then it is *directly correlated*. If, on the other hand, the lower the value of an attribute the more likely the web page is of a particular genre, then it is *inversely correlated*. Otherwise, the attribute is not correlated with the genre. If the correlation between an attribute and the likelihood of a particular appropriateness genre is near zero (or less than 0.18 in our case, which is determined experimentally), then this attribute is considered not to be correlated with that genre.

In formulating the likelihood value that a web page is of a particular appropriateness genre, the value of each (directly or inversely correlated) attribute is normalized. The normalized values are used as input to a SVM for model training on each appropriateness genre. Normalizing the value before applying a SVM is very important. The reason is that it avoids attributes with greater numeric ranges dominating those with smaller numeric ranges. This normalization includes classifying their values into four categories by means of the corresponding cut-off values and then discretizing these categories into their normalized values. For a particular attribute *i*, its three cut-off values for the four categories are $A_i^{0.2}, A_i^{0.5}$ and $A_i^{0.8}$, respectively, as discussed above. With these cut-off values, the normalized value of each attribute *i* (A_{norm_i}) is defined as follows:

$$A_{norm\_i} = \begin{cases} 1 & if \quad A_i \geq A_i^{0.8} \\ \frac{2}{3} & if \quad A_i^{0.8} > A_i \geq A_i^{0.5} \\ \frac{1}{3} & if \quad A_i^{0.5} > A_i \geq A_i^{0.2} \\ 0 & otherwise \end{cases}$$

Equation 1:
Directly correlated attribute normalization.

$$A_{norm\_i} = \begin{cases} 1 & if \quad A_i \leq A_i^{0.2} \\ \frac{2}{3} & if \quad A_i^{0.2} < A_i \leq A_i^{0.5} \\ \frac{1}{3} & if \quad A_i^{0.5} < A_i \leq A_i^{0.8} \\ 0 & otherwise \end{cases}$$

Equation 2:
Inversely correlated attribute normalization.

4.4 Appropriateness Metrics Implementation

As the implementation of the three metrics for classifying the appropriateness genre of a web page is very similar, only the implementation of the *scholarly* metric will be discussed in this section. The details of the implementation of the other two metrics, namely, the *news/general interest* metric and *popular* metric can be found in [16].

To understand the characteristics and the usefulness of the attributes for classifying scholarly pages, a sample data set of more than 125 web pages from different conferences (such as the WWW and META conferences) was analysed. A total of 41 attributes were measured of which those shown in Table 4 were found, experimentally, to be useful in determining whether a web page was a scholarly web page. The results of the measurements are also shown in Table 5 and the distribution of the values of the various attributes is shown in Figure 5 in the Appendix. (Table 6 and Table 7, and Figure 6 and Figure 7 in the Appendix show the measurement and distribution of the values, respectively, for the

attributes that were found to be useful in determining whether a web page was a news/general interest or popular web page.)

5. Experimental Results

In our experiments, the data sets were generated from the ranking results of the Google search engine. They were composed of the first 100 pages returned from Google for the queries "data quality"[c] and "computer games". The contents of these pages were analysed with respect to different web page attributes. The web pages in the data sets were classified manually to determine their appropriateness genre (i.e., scholarly, news/general interest or popular). Within the data sets, "noisy" web pages were first removed, such as pages that were inaccessible (as reflected by their HTTP status code). Finally, 94 web pages were left in both data sets.

To show the effectiveness of the three proposed metrics in classifying web pages according to their appropriateness genre, all the pages in the data set were ranked using each metric. The resulting ranking for the first x pages was then compared with the original ranking by Google as well as with the expected value for the genre of web page under consideration. The expected value is the number of web pages of the genre under consideration that should appear in the first x pages given the distribution of that genre of web page in the entire population[d]. It is calculated by multiplying the ratio of web pages of a particular genre (e.g., scholarly) in the data set with the number x. Hence, for a particular genre, if, in the first x pages, the number of those pages ranked by its corresponding metric is larger than that ranked by Google or given by the expected value, then this shows that the metric can effectively identify pages of that genre.

According to a study and analysis of user's queries on the web [12], about 80% of users will not view more than two query result pages. Usually, each page contains ten query results and hence x was set to 20 in our experiments.

In the data set "data quality", 38 (~40%), 46 (~49%) and 10 (~10%) of the 94 pages were *scholarly* pages, *news/general interest* pages and *popular* pages, respectively (see Table 1). When ranked using the metric that identifies scholarly pages, the first 20 pages contained 16 scholarly web pages (i.e., 80% of the first 20 pages were scholarly pages and ~42% of the scholarly pages were ranked in the first 20 pages) (see Table 1 and Table 2). The original ranking from Google contained only 7 scholarly pages in the first 20 pages, while the expected value of scholarly pages in the first 20 pages is 8.08 (i.e., $\frac{38}{94} \times 20$). When ranked using the metric that identifies news/general interest pages, the first 20 pages contained 13 news/general interest pages. The original ranking from Google contained 12 such pages in the first 20 pages, while the expected number of news/general interest pages in the first 20 pages is 9.79 (i.e., $\frac{46}{94} \times 20$). Finally, when ranked using the metric that identifies popular pages, the first 20 pages contained 5 popular web pages. The original ranking from Google did not contain any popular pages in the first 20 pages, while the expected number of popular pages in the first 20 pages is 2.128 (i.e., $\frac{10}{94} \times 20$).

---

[c] This relatively academic query is selected so as to show the effectiveness of the three metrics more clearly. We have also experimented with other queries [16].

[d] We assume a random distribution of the genre of web page under consideration in the population.

Table 1. Ranking results of the three web metrics.

Web page genre (X)	**Scholarly**	**News / General Interest**	**Popular**
In the first 20 pages:			
Ranked by its metric	16	13	5
Ranked by Google	7	12	0
Expected value	8.08	9.79	2.128
Total no. of pages:	94		
Total number of X pages	38	46	10
Percentage of total pages	~40%	~49%	~10%

Table 2. Percentages of different genres of web pages ranked using their corresponding metrics.

Web page genre (X)	**Scholarly**	**News / General Interest**	**Popular**
% of the first 20 pages ranked by metric X that were of genre X	80%	65%	25%
% of the first 20 pages ranked by Google that were of genre X	35%	60%	0%
% of the first 20 pages expected to be of genre X	~40%	~49%	~10%
% of the genre X pages that were ranked in the first 20 pages	~42%	~28%	50%

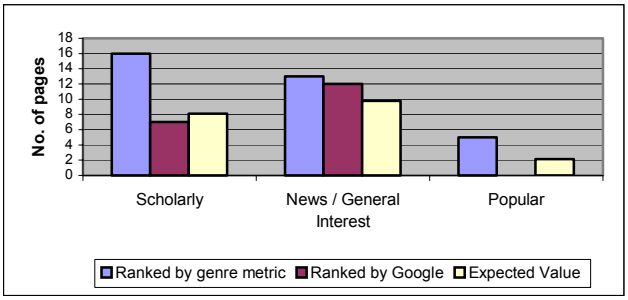

Figure 3. Effectiveness of the three metrics for the “data quality” data set..

Figure 3 clearly shows that the three metrics are effective at ranking a web page of their own genre. However, the effectiveness of the metric that identifies news/general interest pages is not as high as the other metrics. This may be due to the fact that Google’s ranking is biased toward finding news/general interest pages or perhaps it is due to the fact that there are so many news/general interest pages in the population (~50%), it is not hard to get a large percentage of them in the first 20 pages using Google. Similar experimental results were obtained for the data set "computer games" as shown in Figure 4.

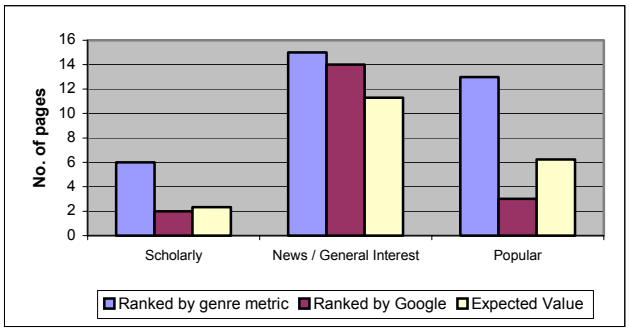

Figure 4. Effectiveness of the three metrics for the "computer games" data set.

In summary, the experimental results show that the three metrics developed to classify the appropriateness genre of web pages (i.e., scholarly, news/general interest, and popular) can model and identify the corresponding genre correctly.

6. Conclusions and Future Work

Currently, primarily link-based metrics are used in most search engines to rank the search results. Most of the results returned are relevant to a query. Sometimes, however, query relevance is insufficient to satisfy a user's task and information needs. It may also be important to return pages that are *appropriate* for the user.

To understand this difference, a framework for web data quality is proposed and six additional dimensions, *navigation, visual appearance, appropriateness*, *cohesiveness, minimality* and *popularity,* related to web data quality are introduced. With reference to this framework, it is found that link-based metrics are focused on only one or two dimensions of web data quality (i.e., *usefulness* and *believability*). In order to have a more complete picture of the factors required to produce a good quality ranking of search results for users, search engines should consider more web data-quality dimensions.

Accordingly, the new web data-quality dimension, *appropriateness*, is further explored with respect to its measurement and use in ranking search results. To model the *appropriateness* of a page for a query, web pages are classified into three main genres according to their linguistic and visual complexity: *scholarly, news/general interest* or *popular*. To measure the likelihood that a page is one of the above three genres, various attributes of a web page are measured quantitatively. This new dimension of web data quality, *appropriateness,* is then computed by three new metrics using a SVM to classify web pages according to their appropriateness genre. From the experiments, the effectiveness of the metrics in classifying web pages according to their appropriateness genre has been shown.

With the introduction of the appropriateness web data-quality dimension, a new way to rank search results is possible. When searching the Web, users can enter both a query and their expected appropriateness genre to a search engine. The metrics can then allow the search engine to rank a web page not only by its relevance to the user's query, but also by its appropriateness to their task and information needs.

One possible concern regarding the proposed metrics for classifying the appropriateness genre of web pages is that they may be too time-consuming to compute. However, as the measurement of a web page's attributes and the computation of a web page's appropriateness genre can be done in the pre-processing indexing stage, it can be done off-line and the appropriateness genre can be appended to the

existing indexes of a search engine for use when answering user queries. Hence, there should be no significant increase in query runtime.

To further enhance the usefulness of the appropriateness dimension, automatic tracking of user preference for the genre they would like returned would definitely be useful. With such tracking, it would be possible to develop a user preference profile. The preferred genre of web page for a particular query could be kept in this profile allowing a search engine to automatically know the preferred genre for a query and, thus, return the most appropriate pages for that query.

Finally, considering the appropriateness dimension to rank search results is a first step to improve the quality of pages returned for a query. The measurement and use of the other new web data-quality dimensions is currently under investigation.

References

1. J.E. Alexander, M.A. Tate. *Web Wisdom: How to Evaluate and Create Information Quality on the Web*, Lawrence Erlbaum Associates Inc., 1999.
2. B. Amento, L. Terveen and W. Hill. "Does "Authority" mean quality? Predicting expert quality ratings of web documents," *Proc. 23rd ACM SIGIR Conf.*, 296-303, 2000.
3. R. Baeza-Yates, F. Saint-Jean and C. Castillo. "Web structure, dynamics and page quality," *Proc. SPIRE 2002*, LNCS, Springer, 2002.
4. S. Brin and L. Page. "The anatomy of a large-scale hypertextual web search engine, *Proc. 7th World Wide Web Conf.*, 107-117, 1998.
5. California Medical Association. *How to Evaluate Medical Information Found on the Internet.* http://new.cmanet.org/publicdoc.cfm/60/0/GENER/99
6. J. Cho and S. Roy. "Impact of web search engines on page popularity," *Proc. 13th World Wide Web Conf.*, 20-29, May 2004.
7. Cornell University Library. *Distinguishing Scholarly Journals from Other Periodicals.* http://www.library.cornell.edu/okuref/research/skill20.html
8. K. Crowston and M. Williams. "Reproduced and emergent genres of communication on the Word-Wide Web," *Proc. Thirtieth Annual Hawaii Intl. Conf. on System Sciences*, Vol. 6, 30-39, 1997.
9. C. Fox, A. Levitin and T. Redman. "The notion of data and its quality dimensions," *Information Processing and Management* **30**(1), 9-19, 1994.
10. J. D. Graofalakis, P. Kappos and D. Mourloukos. "Web site optimization using page popularity," *IEEE Internet Computing* **3**(4), 22-29, 1999.
11. R. Gunning. *Techniques of Clear Writing, revised edition*. McGraw-Hill, New York, 1968.
12. B.J. Jansen, A. Spink and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web," *Information Processing and Management* **36**(2), 207–227, 2000.
13. J. Kleinberg. "Authoritative sources in a hyperlinked environment," Proc. 9th Symp. on Discrete Algorithms, 668-677, 1998.
14. J.T. Kwok. "Automated text categorization using support vector machines," *Proc. Intl. Conf. on Neural Information Processing*, 347-351, 1998.
15. L.L. Pipino, Y.W. Lee, and R.Y. Yang. "Data quality assessment," *Communications of the ACM* **45**(4), 211-218, 2002.
16. J. C.C. Pun and F.H. Lochovsky. *Finding an Appropriate Web Page.* Technical Report, HKUST-CS-04-05, 2004.
17. D. Roussinov, K. Crowston, M. Nilan, B. Kwasnik, J. Cai and X. Liu. "Genre based navigation on the Web," *Proc. Thirty-Fourth Annual Hawaii Intl. Conf. on System Sciences*, Vol. 10, 2001.
18. M. Shepherd and C. Watters. "Identifying web genre: hitting a moving target," *WWW 2004 Conf. Workshop on Measuring Web Search Effectiveness: The User Perspective*, 2004.
19. R. Song, H. Liu, J.R. Wen and W.Y. Ma. "Learning block importance models for web pages," *Proc. 13th World Wide Web Conference*, 203-211, 2004.

20. D.M. Strong, Y.W. Lee and R.Y. Wang. "Data quality in context," *Communications of the ACM* **40**(5), 103-110, 1997.
21. UC Berkeley Library. *Evaluating Web Pages: Techniques to Apply & Questions to Ask.* http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Evaluate.html
22. V. Vapnik. The Nature of Statistical Learning Theory. Spring Verlag, 1995.
23. C.D. Waddell. *Applying the ADA to the Internet: A Web Accessibility Standard.* http://www.rit.edu/~easi/law/weblaw1.htm
24. J. Wang and F.H. Lochovsky. "Data-rich section extraction from HTML pages," *Proc. 3rd Intl. Conf. on Web Information System Engineering,* 313-322, 2002.
25. *The Webby Awards Categories.* http://www.webbyawards.com/main/webby_awards/index.html#categories
26. M.A. Winker, A. Flanagin, B. Chi-Lum, J. White, K. Andrews, R.L. Kennett, C.D. DeAngelis and R.A. Musacchio. *Guidelines for Medical and Health Information Sites on the Internet: Principles Governing AMA Web Sites*. http://www.ama-assn.org/ama/pub/category/1905.html
27. J.C. Wyatt. "Commentary: measuring quality and impact of the World Wide Web," *British Medical Journal*. **314**(7098), 1879-1880, 1997.
28. R.Y. Yang, M.P. Reddy and H.B. Kon. "Toward quality data: an attributed-based approach," *Decision Support Systems* **13**(3), 349-372, 1995.
29. P. Zhang and G.M. von Dran. "User expectations and rankings of quality factors in different web site domains," *Intl. J. of Electronic Commerce* **6**(2), 9-33, Winter 2001-2002.

Appendix

Table 3. Dimensionality of Data Quality -- How Applicable to Web Data?

DQ Dimensions	**DQ Category [20]**				**Descriptions [15, 20, 28]**	**Applicable to Web?**
	Intrinsic	Contextual	Accessible	Representational		
Accessibility			X		User must be able to get the data and user has the means to get the data.	Web accessibility relates to the information on a web page that can be accessed by the broadest range of users of computers and communications equipment, regardless of age or disability [23].
Availability					Exists in some form that can be accessed; percentage of time an information source is "up".	Web pages can be publicly accessible. It has also been interpreted as the number of broken links contained in the web pages.
Access security			X		Relates to the confidential nature of data which require special access permission (e.g. sign on, enter password).	Intranet pages can be accessed by registered users only.
Interpretability (or Ease of Understanding)				X	User understands the syntax and semantics of the data, and can interpret values correctly; degree to which the information conforms to technical ability of the consumer; the extent to which data is in appropriate languages, symbols and units and the definitions are clear.	Free form of web pages makes it difficult to evaluate the interpretability.
Syntax						Use of semi-structure (e.g. table, point form) and special HTML tags (e.g. titles, headings, emphasized words) can help user interpret the content in a web page.
Semantics						Not easy to interpret unless the web page has been expressed in XML.
Usefulness					Data can be used as an input to the user's decision-making process.	
Relevance		X			Fits requirement for making the decision; degree to which information satisfy users need; the extent to which data is applicable and helpful for the task at hand; information that directs to the point; having to do with the matter at hand.	Relates to the measurement of relevance between queries and search results.

Timeliness (or Freshness)		X			Whether the recorded data value is not out of date and availability of output on time; the extent to which the data is sufficiently up-to-date for the task at hand; the time difference between when the process is supposed to have created a value and when it actually has.	Some web pages are rather static and their contents do not change frequently. Frequently updated web pages can also be found in financial and news types of web sites.
Currency					Degree to which data in question is up-to-date; when the data item was stored in the database.	Dates are not always deliberately stated in a web page. Usually, the only date information can be obtained from the date of the web page (or file). However, whether it means date first created, last updated, or placed on the web is not clear.
Non-volatile					How long an item remains valid.	
Believability	X				User can use the data as a decision input; the extent to which data is regarded as true and credible.	
Completeness		X			The extent to which data is not missing, is of sufficient breadth and depth for the task at hand and values are present in a data collection; all values for a certain variable are recorded.	Difficult to define all values as the scope of the Web is unlimited.
Consistency				X	It can refer to several aspects of data: *values of data*, *logical representation of data* and *physical representation of data*. Relating consistency to the values of data, a data value is expected to be the same in all cases.	As web page is free format; consistency of a web page (and its content) is difficult to guarantee.
Credibility (or Reputation or Authority)	X				Degree to which the information or its source is in high standing, authoritative and high reputation. The extent to which data is highly regarded in terms of its source or content; information dependable.	It relates to the reputation of an organization that owns the web page. It can be estimated by the hub/authority rank or by any external recognition of the web site (e.g. awards, no. of visited times).
Accuracy (or Correctness)	X				The recorded value is in conformity with the actual value; ratio of number of correct values in a source to the overall number of values in a source; information has no error, correct, exact, precise, right and true.	Almost everyone can publish on the Web and hence the content of web pages may not be verified by editors. Difficult to guarantee the accuracy of content.
Objectivity	X				The extent to which data is unbiased, unprejudiced, and impartial.	Difficult to judge as the goals or aims of the persons presenting the material often are not clearly stated.

Table 4. Web page attributes used to determine appropriateness.

Attribute	**Description**
$words_{FULL}$	Total no. of words on a page
$words_{DSE}$	Total no. of words on a page after DSE [24] extraction
Hard words count	Total no. of hard words on a page
Numeric words count	Total no. of numeric words on a page
Sentences count	Total no. of sentences on a page
Short form count	Total no. of short form (including abbreviations) on a page
Point form count	Total no. of point form used on a page
Font face count	Total no. of font face used on a page
Font size count	Total no. of font size used on a page
Colour count	Total no. of colour used (font and background) on a page
Images count	Total no. of images on a page
Useful images count	Total no. of useful images (i.e., images with size > θ) on a page
Ratio of useful images	Useful images count / Images count
Unique links count	Total no. of unique links on a page
Anchor text count	Total no. of anchor text on a page
Ratio of anchor text	Anchor text count / $words_{DSE}$
HTTP status code	HTTP status code returned from a page
Minimality	Proportional of useful information of a page
Fog index	Gunning Fog Index [11] is used to measure the *readability* of a document
Ratio of numeric words	Numeric count / $words_{DSE}$
Ratio of short forms	Short form count / $words_{DSE}$
Ratio of point forms	Point form count / Sentences count
Ratio of hard words	Hard words / $words_{DSE}$

Table 5. Web page attribute characteristics of scholarly pages.

Attribute (A_i)	**Relationship**	**Average ($\overline{A_i}$)**	**Top 20% ($A_i^{0.2}$)**	**Top 50% ($A_i^{0.5}$)**	**Top 80% ($A_i^{0.8}$)**
Fog Index	+	14.491	16	14.4	12.8
Minimality	+	0.986	1	1	1
Numeric words count	+	112.63	162	96	34
Ratio of numeric words	+	0.0172	0.0240	0.0131	0.0084
Short forms count	+	96.45	143	80	43
Ratio of short forms	+	0.0174	0.0253	0.0149	0.0085
$\text{Words}_{\text{DSE}}$	+	6123.75	8694	6072	4198
Font face count	-	1.75	3	1	1
Font size count	-	1.98	3	2	1
Colour count	-	1.18	1	1	1
Point forms count	+	21.91	35	14	2
Ratio of point forms	+	0.0655	0.1018	0.0379	0.0051
No. of hard words	+	1119.58	1475	1124	734
Ratio of hard words	+	0.1905	0.2167	0.1863	0.1553

(+) means directly correlated, and (-) means inversely correlated

Sources of the scholarly pages: WWW and META conferences.

Table 6. Web page attribute characteristics of news / general interest pages.

Attribute (A_i)	**Relationship**	**Average ($\overline{A_i}$)**	**Top 20% ($A_i^{0.2}$)**	**Top 50% ($A_i^{0.5}$)**	**Top 80% ($A_i^{0.8}$)**
Useful images count	+	2.91	4	3	2
Ratio of useful images	+	0.105	0.2	0.1176	0.0769
Fog index	-	12.274	14.4	12.4	10.4
Minimality	-	0.913	1	0.96	0.8
Numeric words count	-	17.94	23	14	6
Short forms count	-	15.45	16	12	7
Ratio of short forms	-	0.0111	0.0203	0.0106	0.0048
$\text{Words}_{\text{DSE}}$	-	2028.83	2264	1170	624
Font face count	+	1.2	2	1	1
Font size count	+	1.2	2	1	1
Colour count	+	7.17	9	9	4
Hard words count	-	302.03	382	185	130
Unique links count	+	40.686	47	43	32
Ratio of anchor text	+	0.097	0.1336	0.0667	0.0344

(+) means directly correlated, and (-) means inversely correlated

Sources of the news/general interest pages: National Geographic, Scientific American and The Atlantic Monthly.

Table 7. Web page attribute characteristics of popular pages.

Attribute (A_i)	**Relationship**	**Average ($\overline{A_i}$)**	**Top 20% ($A_i^{0.2}$)**	**Top 50% ($A_i^{0.5}$)**	**Top 80% ($A_i^{0.8}$)**
Images count	+	33.26	47	35	25
Useful images count	+	5.32	8	5	3
Ratio of useful images	+	0.126	0.2647	0.1389	0.1111
Fog index	-	12.068	14	11.2	10
Minimality	-	0.726	1	0.8	0.48
Numeric words count	-	11.79	18	7	3
Ratio of numeric words	-	0.016	0.0244	0.0112	0.0065
Short forms count	-	19.58	21	15	2
Ratio of short forms	-	0.0279	0.0545	0.0146	0.0046
$Words_{DSE}$	-	830.49	1047	521	287
Font face count	+	1.13	1	1	1
Font size count	+	1.13	1	1	1
Colour count	+	9.91	17	9	5
Hard words count	-	108.17	136	66	46
Ratio of hard words	-	0.1418	0.1652	0.1429	0.1011

(+) means directly correlated, and (-) means inversely correlated

Sources of the popular pages: Time, Sports Illustrated, Readers Digested and Vogue.

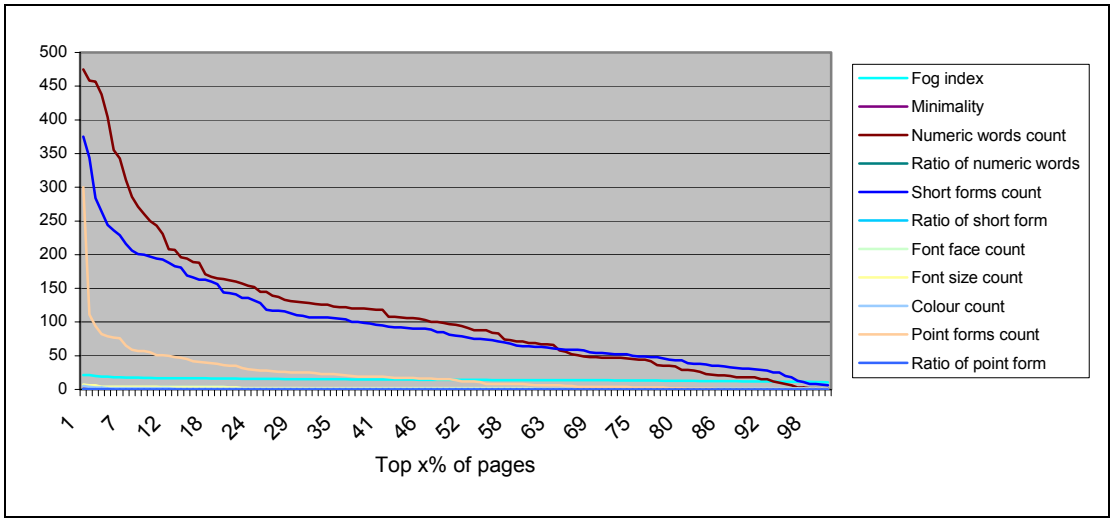

Figure 5. Distribution of the values of different web page attributes for scholarly pages.

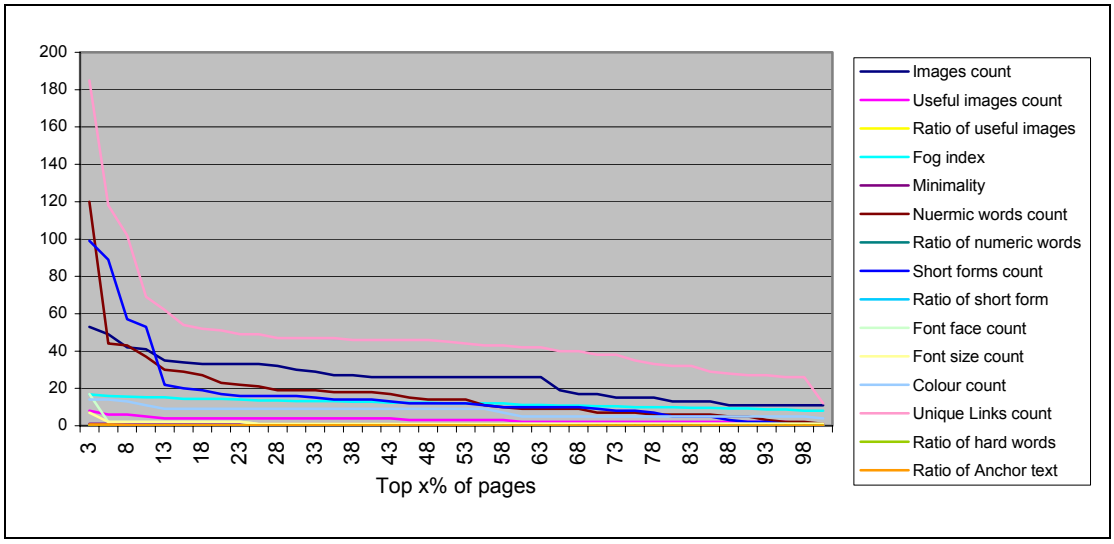

Figure 6. Distribution of the values of different web page attributes for news/general interest pages.

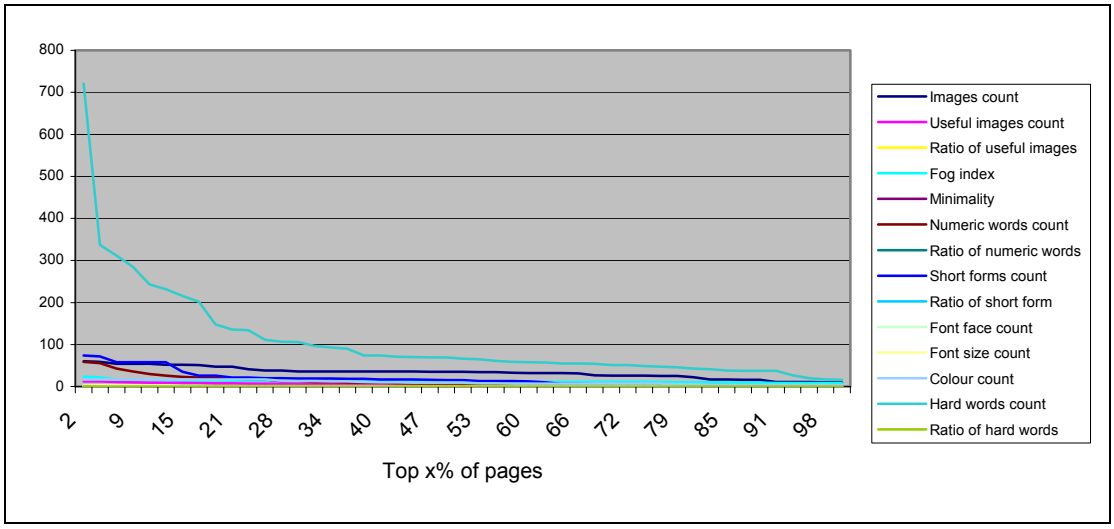

Figure 7. Distribution of the values of different web page attributes for popular pages.