INITIAL STUDY OF A "QUICK AND DIRTY" WEBSITE DATA QUALITY INDEX

(Research-in-progress)

Irit Askira Gelman
University of Arizona
askirai@email.arizona.edu
Anthony L. Barletta
University of Arizona
albarletta@msn.com

In this paper we report on a preliminary test of a potentially useful, "quick and dirty" indicator for assessing the data quality of web pages, websites, and web domains. An underlying perception which motivates this work is that the spelling error rate of a document can serve as a rudimentary proxy for the degree of quality control exercised in its creation, and, subsequently, indicate its data quality. We examine the reliability and validity of an error index that utilizes the reported hit counts of search engine queries on a set of common spelling errors in the English language. While our tests show positive results, the proposed approach requires further research.

1. INTRODUCTION

As of June 2008, the number of websites on the Internet has been estimated to be over 100 million[1], and these websites contain over 60 billion web pages[2]. This huge information store is, in essence, a democratic environment. In order to publish a web page, one does not have to go through a publisher or any other institution that controls the content or format of the information. The emergence of free web hosting services has facilitated the publication of a variety of materials and has led to the creation of countless personal and social networking pages, as well as sites by small organizations or interest groups, families, and others. Obviously, the quality of the information on the web is very heterogeneous.

Search engines alleviate the task of identifying relevant, high quality information on the web. At the core of a search engine like Google there is a weighted voting mechanism that uses the link structure of the web as an indicator of an individual page's value. Google's PageRank and similar algorithms that exploit the wisdom of crowds[3] exhibit impressive performance overall (e.g., [2]). Nonetheless, the use of a search engine leaves significant uncertainty regarding the quality of its

[1] http://www.domaintools.com/internet-statistics/ , accessed on June 4 2008.

[2] http://www.worldwidewebsize.com/ , accessed on June 4 2008.

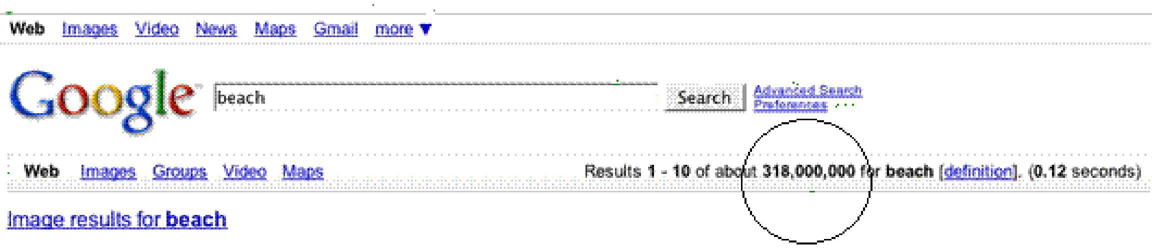
[3] James Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economics, Societies, and Nations*. Doubleday, 2004.

ranking (e.g., [24]). Nor does it resolve the data quality concerns of other applications on the Internet.

While educational institutions are trying to sensitize students to the great variation in the reliability and credibility of data on the web and equip them with basic guidelines for assessing information (e.g., [21],[26],[27]), there is a growing number of studies that focus on the assessment of the quality of web data under miscellaneous assumptions. A broad framework for the assessment of web data is proposed by Pun and Lochovsky [17]. This framework examines four data quality dimensions: accessibility, interpretability, usefulness, and believability. Pun and Lochovsky [18] develop an algorithm that identifies high quality web pages based on an analysis of their cohesiveness. Their work is founded on the understanding that high quality documents are very cohesive. A few models that target certain subsets of the web have been proposed as well. Eppler et al. [8] test a conceptual framework that specifies four information quality dimensions for a content-driven website (e.g., news portals): relevance, soundness, process, and infrastructure. Stvilia, Gasser, and Smith [23] demonstrate the systematic development of an information quality model for Wikipedia, the free encyclopedia. Cappielo and Pernici [5] describe a methodology for the identification and correction of data quality problems in self-healing web services.

This paper adds to the literature on web data quality assessment. The underlying perception that motivates this study is that the spelling error rate of a web document can serve as a rudimentary proxy for the degree of quality control exercised in its creation. Consequently, the spelling error rate signals the quality of the document in general. Basic guidelines that are suggested by educational institutions for the assessment of web data advise students to look for spelling errors as a sign of low information credibility and quality (e.g., [21],[26],[27]). Our own anecdotal experience has indicated the prevalence of spelling errors in social forum exchanges, personal websites, informal applications of the wiki software, etc. Errors can also be quite common in immediate press releases and other texts that are characterized by weak quality control.

In this paper we report on an initial test of a simple, potentially useful, "quick and dirty" index for assisting in the evaluation of website information quality. To the extent that there is a link between spelling error rates and data quality in general, this index is aimed at taking advantage of that link. In particular, we examine the reliability and validity of an indicator that utilizes the reported hit counts of search engine queries on a small set of common spelling errors (Figure 1 shows an example of Google's hit count report). Notably, this metric can be easily automated.



Results 1 - 10 of about 318,000,000 for beach [definition]. (0.12 seconds)

Figure 1: Google's hit count (circled).

Recently there has been a growing interest in the information that search engines supply and its potential for providing new solutions to vital problems. The hit count has lately been applied in a distance measure for automatic word meaning discovery [6]. Pion and Hamel [16] investigated the hit count as an alternative for prediction markets. Krebs [12], and Simkin and Roychowdhury [22], studied fame, and Bagrow et al. [3] explored the correlation between fame and merit. In this context, our contribution lies in the new application that we propose for the hit count and the initial validation that is reported in this paper. We believe that search engine outputs can benefit data quality research and practice in a variety of ways. We hope that this work will draw attention for search engines as an interesting potential for future work.

The use of the hit count is controversial, however. The (public's) understanding of the hit count is too limited to determine its usefulness for our purposes without inquiry. Therefore, we explore the validity of the proposed index with special emphasis on its use of the hit count. We also examine the validity of this indicator in view of theory-based constraints on error rates, and the expected link with the degree of quality control.

The structure of this paper is as follows: Section 2 offers a short overview of the objectives of this research and anticipated challenges. Section 3 describes the method of the initial inquiry. Section 4 presents our results. Section 5 concludes this paper with a discussion of the findings and future research directions.

2. RESEARCH OBJECTIVES AND CHALLENGES

The goal of this research is to develop a data quality indicator which is based on the reported hit counts of search engine queries on a set of common spelling errors. At present we focus on documents in the English language and employ Google's hit count.

The development of the desired metric involves several challenges. First, instead of an exhaustive spelling error test, we aim at a simple method that tests a minimal number of spelling errors. Clearly, the number of spelling errors and their selection should be determined carefully. To enable a valid assessment of documents, a good selection would consist of spelling errors that are common enough in the target document population.

A second set of issues is related to the intended utilization of a search engine's hit count. A hit count "indicates the total number of results" [9]. In other words, the hit count is an estimate, not a precise number. The use of a hit count became popular in recent years for both practical and research purposes; a few examples of research works that apply the hit count have been mentioned in the introduction. A major advantage of the hit count is its cheap entry cost. However, there is substantial controversy regarding the validity and usefulness of solutions that utilize it. The conclusion of a recent paper in the journal "Computational Linguistics" [11] is unequivocally summarized by its title: ”Googleology is Bad Science.” A Wikipedia article on the use of a search engine's output [29] concludes that "Depending on the subject matter, and how carefully it is used, a search engine test can be very effective and helpful, or produce misleading or non-useful results. In most cases, a search engine test is a first-pass heuristic or 'rule of thumb'."

Specific weaknesses of the hit count that can affect its usefulness for our purpose include (e.g., [14]):

1. The hit count might fluctuate in a short time because of the constant update of distributed indices.

2. The search is not context-sensitive, and the hit count is affected accordingly. For instance, a search for the common misspelling "grammer" would also return results containing information about a person whose family name is Grammer.

3. The results obtained using Boolean operators such as OR and AND sometimes violate the basic set theory laws. For instance, the hit count of “A and B” can be higher than the hit count of “A.” Recent studies have proposed methods for improving the accuracy of the hit count [14].

In addition, the hit count designates the *number of pages* that contain a specified data item rather than the *number of times* that the item appears. While this property is not necessarily a weakness of the hit count, it may detract from the value of our metric.

Clearly, the reliability and validity of a metric that applies the hit count are uncertain and cannot be taken for granted.

Notably, there is an obvious need to explore the fundamental assumption that the spelling error rate is related to web document quality. The hypothesized link between spelling errors and *quality control* implies that we may not necessarily find a match between spelling errors and a specific data quality dimension such as accuracy, completeness, or the like [28]. According to this logic, the types of data quality deficiencies that may be observed in connection with spelling errors can vary across different documents. Therefore, validation of the hypothesized link will be based on an aggregate measure that would account for multiple quality dimensions.

3. METHOD

3.1 Spelling errors

In the early 1990s, a Harvard University cataloguer named Jeffrey Beall proposed the “Dirty Database Test” [1]. This test incorporates a list of ten English spelling errors for studying the error rate in library catalogue databases. Variations of the Dirty Database Test are popular in data quality studies of library databases (e.g., [4],[19]). To a large extent, Beall’s test targets spelling errors that professional librarians commit in the data entry process. For this initial inquiry, however, we have sought a word selection that would apply equally well to varied texts, created through diverse professional, administrative, business, academic, or leisure activities. Therefore, we have preferred spelling errors that are universal, rather than the choice of a small community of professionals. For the sake of simplicity, however, our spelling error set in this preliminary study is of the same size as Beall’s list.

Another pre-requisite of a good selection given the properties of current search engines is that it should account for the lack of context-sensitivity of the search engine.

Our set consists of ten spelling errors that have been recognized by Microsoft as pervasive errors.[4] Specifically, this choice consists of an arbitrary selection of ten spelling errors from the English version of Microsoft Word’s AutoCorrect function (Table 1). The AutoCorrect feature utilizes lists of common misspellings. Versions of these lists exist for various languages. A Microsoft representative explained to us that “There is one 'default' set of AutoCorrects for English that is used to create the starting position for English (UK), English (Aus), etc…There may be a few new entries added each release but only a few in recent versions and not enough to lose sleep over.”

Subsequently, our word selection is based on the default file that is shared by all the English speaking countries.

[4] One misspelling in our list (receive) coincides with Beall’s list.

	Spelling Error	Correct Spelling
1	Recieve	Receive
2	Accomodate	Accommodate
3	Accross	Across
4	Truely	Truly
5	Acheive	Achieve
6	Affraid	Afraid
7	Agressive	Aggressive
8	Appearence	Appearance
9	Tomorow	Tomorrow
10	Arguement	Argument

Table 1: Selected spelling errors and the matching correct spellings.

Top-Level Domain	Type	Country / intended use	English official Language?
.gov	sponsored	Reserved exclusively for the United States government	Yes
.edu	sponsored	Reserved for post-secondary institutions accredited by an agency on the U.S. Department of Education's list of Nationally Recognized Accrediting Agencies	Yes
.com	generic	Unrestricted	
.org	generic	Unrestricted	
.info	generic	Unrestricted	
.aero	Sponsored	Reserved for members of the air-transport industry	
.mil	Sponsored	Reserved exclusively for the U.S. Military	Yes
.jp	country code	Japan	No
.cn	country code	China	No
.it	country code	Italy	No
.fr	country code	France	No
.il	country code	Israel	No
.cl	country code	Chile	No
.gr	country code	Greece	No
.ru	country code	Russian Federation	No
.eg	country code	Egypt	No
.mx	country code	Mexico	No
.au	country code	Australia	Yes
.in	country code	India	Yes
.nz	country code	New Zealand	Yes
.uk	country code	United Kingdom	Yes
.za	country code	South Africa	Yes

Table 2: A list of the top-level domains. [5]

[5] The data about the type and country/intended use are based on IANA's (Internet Assigned Numbers Authority) root zone database (www.iana.org).

3.2 Web document sets

Google enables users to limit the scope of a search through an advanced search option, i.e., the "site" constraint. For example, the search for "receive site:.gov.au" outputs pages that contain the word receive and belong to the domain of the Australian government. Obviously, the matching hit count will agree with the site constraint.

Our study exploits the "site" option of Google:

- One test focuses on the pages of a popular website, namely, en.wikipedia.org, the English component of Wikipedia.
- The remaining tests designate the entire web, as well as 22 top-level domains (TLDs), plus, for each country-code TLD (ccTLD), a second level domain (SLD) of the government of the country (see Table 2 and Table 3).

The selection of the TLDs has been partly arbitrary. However, we have gravitated towards larger domains and included both sponsored and generic TLDs (see TLD type and intended use in Table 2). Also, our initial list encompasses both countries in which English is an official language and countries in which English is not an official language. This preference, just like the decision to include both sponsored and generic TLD's, enables us to explore potential differences between domain classes.

3.3 Spelling Error Index

Given a document set of interest and a spelling error, the proposed index is a value in the interval [0,1] that is calculated according to the following formula (1). Let e_j, *j*=1,..,10, denote the *j*th spelling error in Table 1, and c_j will denote the correct spelling that matches e_j; let d denote the interesting document set, then:

$$ErrorIndex(e_j, d) = \frac{HitCount(e_j, d)}{HitCount(e_j, d) + HitCount(c_j, d) + 1} \qquad (1)$$

According to (1), the index is calculated as the hit count of the spelling error on the assigned document set divided by the sum of the hit counts of the spelling error and the corresponding correct spelling plus one (we add one to avoid division by zero when the hit counts are zero). This index is an estimate of the fraction of the document set that contains spelling errors. Assuming that each flawed document reflects a quality control failure, this index hints to the degree of quality control in the document set. For instance, if the control, and, therefore, document quality itself, is unevenly distributed such that some documents are poorly controlled and the remaining documents are highly controlled, then this index can give us a clue about the proportion of poorly controlled documents in the set.

By applying (1) on each of the spelling errors in Table 1 we will produce ten different index values for each document set. The method in which these values are reconciled is shown below (2). We define the index of a document set as the *average* of the index values that are obtained for the distinct spelling errors:

$$ErrorIndex(d) = AVERAGE\ \{ErrorIndex(e_j, d): j = 1,..,10\} \qquad (2)$$

Admittedly, since the hit count refers to the number of pages that contain a specified data item, rather than the number of times that the item appears, this index neglects some quantitative aspect of the errors. This information might be missed most when the document set is particularly small.

3.4 Tests

Four tests of the proposed index are described next:

1. A test of its reliability through repeated measurements.

2. A test of its validity by verifying the observed values against a value domain which is predicted by theory (in some research areas this type of validity is termed criterion, or concurrent, validity, e.g., [20]).

3. A second criterion validity test examines the correlation between error index values and selected web domain types.

4. A third validity test examines the correlation between spelling error index values and document types in Wikipedia.

Test-retest reliability. Given the reported fluctuations, estimation of the reliability of the index calls for several cycles of measurements. We have conducted a series of four measurement cycles over a time period of two months between March 25, 2008, and May 26, 2008. This timeframe has been selected in order to avoid, on one hand, the significant development of the web over longer periods of time, and to enable, on the other hand, identification of fluctuations that occur over time periods of days or weeks. In each cycle, the error indices of each of the ten spelling errors were recorded for each of the document sets.

A similarity between correlation coefficients enabled us to use the standardized item alpha (Spearman-Brown formula) [20] for assessing aggregate reliability.

Validity test 1. This test relies on a rich theory regarding the accuracy of groups. This literature, both analytical and empirical, has examined democracy and group accuracy for many generations (e.g., [7],[10],[13]). Assumably, our chosen problem area (common spelling errors) does not require exceptional expertise in order to make the correct choice. Subsequently, that theory suggests that the majority of the instances of each word in a diverse environment such as the entire web or significant portions of the web would be spelled correctly. Ordinarily, this understanding would imply also that a majority of the pages should show the correct spelling of a common word, hence, our indices should be lower than 0.5. A high index value will hint to a failure of the underlying hit count, although it cannot rule out the possibility that the "web crowds" are the cause of the unexpected failure, rather than the hit count.

Validity test 2. The second validity test designates an anticipated variation of the proposed index across web domains. Mainly, we focus on a class of websites whose data quality often undergoes a higher scrutiny than many other websites, and uncontrolled information sharing is relatively limited. This class is government websites. In the U.S., the importance of government data quality is indicated by the Federal Data Quality Act of 2001. This law has largely been motivated by the growing dissemination of government information through the Internet. In essence, it is intended to ensure and maximize the quality, objectivity, utility, and integrity of information disseminated by Federal agencies [15]. In recent years, governments around the world are using information technologies to improve information and service delivery. Countries that are leaders in e-government initiatives include Sweden, Canada, the United States, Denmark, Australia, France, the United Kingdom, and Japan [25]. Data quality is an important component of these efforts. Such indirect evidence suggests that government websites would show relatively low error rates.

Consequently, given a set of country-code TLD and corresponding government SLDs, we validate the relationship between the index and domain by testing the null hypothesis that the means of the corresponding indices are equal.

Validity test 3. This test centers on the link between the index and document types in the English version of Wikipedia. Wikipedia documents are organized by categories according to their purpose and use. Most Wikipedia pages belong to one of the categories mentioned below. Interestingly, quality control varies substantially across the categories.

- Wikipedia *articles* are subject to constant supervision and revision; the goal of this collaborative effort is to produce high quality articles.
- *Talk* pages are used for the discussion of content (e.g., articles) and for direct communication between participants. Wikipedia specifies broad standards that editors are expected to follow. In particular, "the policies that apply to articles also apply (if not to the same extent)."[6]
- Wikipedia *project* pages are devoted to the management of a specific topic within Wikipedia as well as the group of editors that collaborate on the topic. Wikipedia offers broad standards that editors are expected to adhere to. However, it recognizes the great variation between different projects, and, furthermore, views group cohesion as a critical asset that should be upheld.[7]
- *User* pages are supported in order to "facilitate communication among participants in its [Wikipedia's] project to build an encyclopedia." By convention, user pages are not edited by others. [8]
- *Special pages* are generated automatically and need no editing. [9]

Evidently, while the quality of articles and special pages is highly scrutinized, documents from the remaining categories are not scrutinized with the same intensity. We have collected data and derived the index values for each of the above document categories. We have divided these values into two subsets based on their quality control characteristics: indices of articles and special pages in one subset, and the remaining index values in the other subset. To verify the link between the proposed index and Wikipedia page types, we test the null hypothesis that the means of the indices in these two subsets are equal.

4. RESULTS

4. 1 Test-retest reliability

In agreement with previous reports, this inquiry has revealed partial irregularity in the behavior of the hit count. Nonetheless, tests of the reliability of the spelling error index have produced remarkably positive results.

The indices of individual spelling errors (1) display substantial variation across different measurements. A variation of up to a ratio of 1:1.5 between different measurements is common. In some cases, we have registered fluctuations of up to 1:10. In these cases, however, one number typically stands out and the remaining values are relatively similar (e.g., the recorded value of the index of "truly" for the web domain .org is 0.148 on the first round, and 0.022, 0.144, and 0.148 on the second, third, and fourth rounds, respectively).

[6] http://en.wikipedia.org/wiki/Wikipedia:Talk_page_guidelines , accessed on September 10, 2008.

[7] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Guide , accessed on September 10, 2008.

[8] http://en.wikipedia.org/wiki/User_page , accessed on September 10, 2008.

[9] http://en.wikipedia.org/wiki/Special:SpecialPages , accessed on September 10, 2008.

	receive	accommodate	across	truly	achieve	afraid	aggressive	appearance	tomorrow	argument
.gov	0.022	0.051	0.006	0.005	0.006	0.001	0.010	0.003	0.000	0.003
.edu	0.150	0.201	0.034	0.059	0.034	0.007	0.042	0.018	0.008	0.023
.com	0.023	0.317	0.127	0.044	0.129	0.115	0.147	0.140	0.137	0.288
.org	0.026	0.315	0.043	0.144	0.021	0.048	0.236	0.018	0.088	0.140
.info	0.110	0.198	0.041	0.082	0.027	0.023	0.137	0.017	0.027	0.021
.aero	0.002	0.024	0.001	0.005	0.001	0.005	0.008	0.004	0.002	0.009
.mil	0.006	0.020	0.001	0.002	0.000	0.000	0.010	0.002	0.000	0.000
.jp	0.069	0.076	0.018	0.049	0.002	0.008	0.074	0.011	0.024	0.010
.go.jp	0.004	0.035	0.001	0.011	0.001	0.000	0.014	0.005	0.000	0.000
.cn	0.098	0.056	0.019	0.135	0.019	0.013	0.084	0.009	0.061	0.016
gov.cn	0.013	0.027	0.004	0.056	0.039	0.010	0.120	0.005	0.034	0.004
.it	0.149	0.263	0.029	0.122	0.015	0.010	0.053	0.004	0.010	0.006
gov.it	0.000	0.106	0.002	0.000	0.000	0.000	0.004	0.018	0.000	0.000
.fr	0.176	0.197	0.044	0.052	0.005	0.546	0.571	0.013	0.003	0.010
gouv.fr	0.001	0.512	0.005	0.001	0.000	0.000	0.936	0.006	0.004	0.000
.il	0.288	0.156	0.020	0.058	0.007	0.009	0.058	0.023	0.026	0.009
gov.il	0.005	0.014	0.001	0.001	0.000	0.002	0.015	0.001	0.001	0.000
.cl	0.011	0.044	0.008	0.029	0.004	0.008	0.122	0.026	0.014	0.000
gov.cl	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
.gr	0.038	0.094	0.014	0.024	0.001	0.009	0.034	0.150	0.012	0.011
gov.gr	0.000	0.032	0.000	0.000	0.014	0.000	0.182	0.000	0.000	0.000
.ru	0.074	0.148	0.029	0.055	0.012	0.020	0.268	0.025	0.013	0.007
gov.ru	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
.eg	0.025	0.103	0.005	0.015	0.004	0.000	0.035	0.007	0.005	0.000
gov.eg	0.003	0.010	0.002	0.007	0.001	0.000	0.035	0.003	0.000	0.003
.mx	0.054	0.106	0.008	0.054	0.009	0.011	0.204	0.022	0.011	0.002
gob.mx	0.010	0.224	0.010	0.002	0.001	0.003	0.053	0.010	0.032	0.000
.au	0.189	0.125	0.084	0.174	0.048	0.028	0.096	0.021	0.045	0.090
gov.au	0.007	0.008	0.004	0.005	0.000	0.004	0.002	0.000	0.000	0.001
.in	0.089	0.088	0.031	0.055	0.014	0.009	0.044	0.030	0.013	0.019
gov.in	0.005	0.028	0.004	0.025	0.002	0.001	0.015	0.005	0.002	0.000
.nz	0.094	0.093	0.043	0.092	0.004	0.008	0.045	0.006	0.022	0.037
govt.nz	0.005	0.007	0.001	0.006	0.000	0.000	0.001	0.001	0.000	0.001
.uk	0.233	0.284	0.231	0.236	0.143	0.055	0.180	0.096	0.071	0.236
gov.uk	0.023	0.025	0.006	0.012	0.003	0.002	0.010	0.001	0.000	0.005
.za	0.046	0.071	0.031	0.072	0.006	0.008	0.026	0.005	0.011	0.016
gov.za	0.002	0.005	0.001	0.008	0.000	0.000	0.010	0.000	0.000	0.000
all web	0.030	0.126	0.009	0.051	0.012	0.016	0.079	0.013	0.011	0.024

Table 3: Index values registered on May 12 2008 (third data collection round)

Despite the fluctuations of the indices of individual spelling errors, the aggregate reliability of the indices of the document sets (2), calculated using the Spearman-Brown formula, is higher than 0.98.

The index values that have been registered on one of the four cycles of measurements that we have conducted are portrayed by Table 3. Due to space limitations, the results of the other three rounds are not shown in this paper.

4.2 Validity test 1

The index satisfies the fundamental requirement that theory establishes. However, the indices of individual spelling errors violate this requirement in four instances. These instances include the words "aggressive" and "afraid" under the ccTLD of France (see, for example, the highlighted values in Table 3), and the words "aggressive" and "accommodate" under the government SLD of France. The index values that have been registered in these unusual cases have violated the fundamental requirements of theory in all four measurement cycles. Because of the uniqueness of these problems and their persistence across all measurements, we have investigated these cases further by reviewing subsets of results. These inspections have indicated that:

- A pair of documents in a government document-base whose descriptions contain the misspelling "accommodate" have been counted over and over again.
- The misspelled version of "aggressive" is the correct spelling of a similar word in the French language.

The misspelling "affraid" is common in English texts in this ccTLD. Forum discussions, newsgroups, personal web pages, blogs, and comparable texts are typical sources of this error. Therefore, possibly, poor knowledge of the web crowds is the cause of this outcome, rather than a shortage of the hit count.

Domain	Index	Domain	Index
go.jp	0.007	.jp	0.035
gov.cn	0.034	.cn	0.114
gov.it	0.014	.it	0.081
gouv.fr	0.15	.fr	0.321
gov.il	0.004	.il	0.078
gov.cl	0.000	.cl	0.025
gov.gr	0.027	.gr	0.043
gov.ru	0.000	.ru	0.079
gov.eg	0.006	.eg	0.023
gob.mx	0.041	.mx	0.054
gov.in	0.009	.in	0.042
govt.nz	0.002	.nz	0.045
gov.za	0.002	.za	0.033
gov.uk	0.009	.uk	0.236
gov.au	0.003	.au	0.103

Table 4: Results of the second validity test.

According to Table 3, the generic TLDs (.com, .info, and .org) have very high error indices, and so do the ccTLDs of the UK and Australia. The TLD of educational institutions in the US, .edu, exhibits a relatively high index as well. A brief investigation of the potential sources of such errors points to student forums, student wiki projects, and other unsupervised web pages. However, overall, sponsored top-level domains demonstrate lower error indices than generic top-level domains. This finding makes sense due to existence of gatekeepers that set rules and restrict the eligibility to use a sponsored TLD. Interestingly, the numbers do not uncover an obvious superiority of ccTLDs of countries in which English is an official language over ccTLDs of countries in which English is not an official language.

4.3 Validity test 2

Table 4 shows the data that we have employed for the calculation of the correlation coefficient. Overall, the index values of government domains are substantially lower than the matching values of the country domains. A two-tailed dependent t-test of the null hypothesis that the means of the two sets are equal shows a p-value of 0.0008. Therefore, the null hypothesis is rejected (α=0.001).

	article	special page	Wikipedia project	talk	user
Recieve	0.002	0.000	0.073	0.030	0.028
Accomodate	0.003	0.000	0.091	0.112	0.149
Accross	0.000	0.000	0.003	0.010	0.012
Truely	0.001	0.000	0.046	0.034	0.049
Acheive	0.001	0.000	0.016	0.009	0.022
Affraid	0.001	0.000	0.000	0.001	0.005
Agressive	0.006	0.000	0.069	0.048	0.053
Appearence	0.001	0.000	0.006	0.007	0.019
Tomorow	0.002	0.000	0.002	0.005	0.004
Arguement	0.003	0.000	0.007	0.032	0.029
Index	**0.002**	**0.000**	**0.031**	**0.029**	**0.037**

Table 5: Results of the third validity test.

4.4 Validity test 3

The index values for each of the interesting document sets are shown in Table 5. For the most part, the index values of Wikipedia project, talk, and user documents are markedly higher than the index values of the articles and special pages. A two-tailed t-test of the null hypothesis that the means of these two sets are equal shows a p-value of 0.00005. Therefore, the null hypothesis is rejected (α=0.001).

5. FUTURE RESEARCH DIRECTIONS

Our tests of the reliability and validity of the error index as a surrogate for the degree of quality control of a given document set have yielded positive results. As stated earlier, however, there is an obvious need to explore the fundamental assumption that the spelling error rate is related to web document quality. The hypothesized link between spelling errors and quality control implies that we may not necessarily find a match between spelling errors and a specific data quality dimension such as accuracy, completeness, or the like. Rather, validation of the hypothesized link should be based on an aggregate measure that would account for multiple quality dimensions.

Our findings point to several additional open issues that should be addressed by future work.

The hit count. Consistent with previous reports, our inquiry demonstrates partial irregularity in the behavior of the hit count. In addition, Google has shown abnormal results for the word "aggressive" in the ccTLD of France, and unreliable counts were shown when interacting with a document-base under .gouv.fr.

The effect of the fluctuations of the hit count is moderated by the definition of the index as an average. That effect may be further reduced by a suitable measurement method that consists of multiple measurements. As explained earlier, we have noticed that sharp fluctuations often generate outliers. These outliers can be discarded before final estimates are produced, such that estimates will be calculated as the average of the remaining values.

The lack of context-sensitivity of the search engine is a problem that calls for caution in the selection of spelling errors (see below). Potential distortions of the hit count as in the case of the French document-base will have to be investigated on a case by case basis.

Spelling errors. The choice of English spelling errors should be further studied and spelling errors in other languages should be considered. While our word selection targets a broad document population, different word sets may better suit different target populations. In this study, especially, the word "aggressive" has been proven to be unsuitable when used in the .fr domain. As for the size of the spelling error list, while the size of our list has been determined in agreement with Beall's suggestion, a larger list size can contribute to the overall performance of the metric. Beall's list has been used for evaluating sizable library databases. A larger spelling error set can increase the sensitivity of our index in small document sets, mainly if the documents are short.

BIBLIOGRAPHY

1. AL Aside, "Ideas: The Dirty Database Test." *American Libraries* Vol. 22, No. 3, March 1991, p. 197.
2. Amento, B., Terveen, L., Hill, W., "Does 'Authority' Mean Quality? Predicting Expert Quality Ratings of Web Documents**"** in *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.
3. Bagrow, J. P. , Rozenfeld, H. D., Bollt, E. M., and ben-Avraham, D. "How famous is a scientist?—Famous to those who know us," *Europhysics Letters*, Vol. 67 , No. 4, 2004, pp. 511-516.
4. Cahn, P. "Testing Database Quality", *Database Magazine*, Vol. 17, No. 1, 1994.
5. Cappielo, C. and Pernici, B., "A Methodology for Information Quality Management in Self-Healing Web Services," *11th International Conference on Information Quality (ICIQ-06)*, MIT, Cambridge MA, 2006.
6. Cilibrasi, R.L., and Vitányi, P.M.B., "The Google Similarity Index," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 3, 2007, pp. 370-383.
7. Nicolas Caritat de Condorcet, Essai sur l'application de l'analyse a la probabilité des décision rendues à la pluralité des voix (Paris, 1785).
8. Eppler, M.J., Algesheimer, R., and Dimpfel, M., "Quality Criteria of Content-driven Websites and their Influence on Customer Satisfaction and Loyalty: An Empirical Test of an Information Quality Framework." *8th International Conference on Information Quality (ICIQ-03)*, MIT, Cambridge MA, 2003.

9. Google Web Search Help Center, “Search Results Page,” http://www.google.com/support/bin/static.py?page=searchguides.html&ctx=results&hl=en [June 5, 2008].

10. Grofman, B., Owen, G., and Feld S.L. (1983) Thirteen theorems in search of the truth, *Theory and Decision*, Vol. 15, No. 3, pp. 261-278.

11. Kilgarriff, A. “Googleology is bad science.” *Computational Linguistics*, Vol. 33 , No. 1, 2007.

12. Krebs, V., “What’s your Google Number?” *International Association for Human Resource Information Management Journa*l, Vol. VII, No. 2, 2003, pp. 40-42.

13. Ladha, K.K., “The Condorcet Jury Theorem, Free Speech, and Correlated Votes,” American Journal of Political Science 36, 1992.

14. Matsuo, Y., Tomobe, H., Nishimura, T. “Robust Estimation of Google Counts for Social Network Extraction.” *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence 2007*, pp. 1395-1401.

15. Office of Management and Budget, “OMB Specific Information Quality Web Page,” http://www.whitehouse.gov/omb/inforeg/info_quality/information_quality.html [June 5, 2008].

16. Pion, S., and Hamel, L.,”The Internet Democracy: A Predictive Model Based on Web Text Mining,” *International Conference on Data Mining*, 2007, pp. 292-300.

17. Pun, J.C.C., and Lochovsky, F.H., “Ranking search results by web quality dimensions,” *Journal of Web Engineering*, Vol. 3, No. 3-4, 2004, pp. 216-235.

18. Pun, J.C.C., and Lochovsky, F.H., “Finding High-Quality Web Pages Using Cohesiveness.” *10th International Conference on Information Quality (ICIQ-05)*, MIT, Cambridge MA, 2005.

19. Randall, B.N., Notes on Operations Spelling Errors in the Database: Shadow or Substance? *Library Resources and Technical Services*, Vol., 43, No. 3, 1999, pp. 161-169

20. Rosenthal, R., and Rosnow, R., *Essentials of Behavioral Research: Methods and Data and Analysis* , McGraw Hill, second edition, 1991.

21. Schrock, Kathy. "Critical Evaluation of a Web Site: Secondary School Level." Kathy Schrock's Guide for Educators, DiscoverySchool.com. http://school.discovery.com/schrockguide/evalhigh.html [June 5, 2008]

22. Simkin, M.V., and Roychowdhury V.P., “Theory of Aces: Fame by Chance or Merit?” *The Journal of Mathematical Sociology* , Vol. 30, No. 1, 2006, pp. 33-41.

23. Stvilia, B., Twidale, M.B., Smith, L.C., Les Gasser “Assessing Information Quality of a Community-Based Encyclopedia.” *10th International Conference on Information Quality (ICIQ-05)*, MIT, Cambridge MA, 2005.

24. Sweetland, J.H., Reviewing the World Wide Web—Theory Versus Reality,” *Library Trends*, Spring 2000.

25. “United Nations E-Government Survey: From E-Government to Connected Governance,” United Nations, Department of Economic and Social Affairs, New York, 2008.

26. Ury, Connie and Lori Mardis. "Evaluating Websites: PART of the Research Process." http://www.nwmissouri.edu/library/courses/evaluation/edeval.htm [5 June 2008]

27. Virginia Tech “Evaluating Internet Information,” http://www.lib.vt.edu/help/instruct/evaluate/evaluating.html [June 5, 2008].

28. Wang, R.Y. and Strong, D.M., “Beyond Accuracy: What Data Quality Means to Data Consumers,” Journal of Management Information Systems Vol. 12, No. 4, 1996.

29. Wikipedia: The Free Encyclopedia, “Wikipedia: Search Engine Test,” http://en.wikipedia.org/wiki/Google_test [June 5, 2008].