Web-based Statistical Fact Checking of Textual Documents

Amr Magdy
Computer and Systems Engineering
Alexandria University
Alexandria, Egypt
amr.magdy@alex.edu.eg

Nayer Wanas
Cairo Microsoft Innovation Lab
306 Corniche El-Nile, Maadi
Cairo, Egypt
nayerw@microsoft.com

ABSTRACT

User generated content has been growing tremendously in recent years. This content reflects the interests and the diversity of online users. In turn, the diversity among internet users is also reflected in the quality of the content being published online. This increases the need to develop means to gauge the support available for content posted online. In this work, we aim to make use of the web-content to calculate a statistical support score for textual documents. In the proposed algorithm, phrases representing key facts are extracted to construct basic elements of the document. Search is used thereon to validate the support available for these elements online, leading to assigning an overall score for each document. Experimental results have shown a difference between the score distribution of factual news data and false facts data. This indicates that the approach seems to be a promising seed for distinguishing different articles based on the content.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods*H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, selection process* H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *web-based services* I.5.1 [**Pattern Recognition**]: Models – *statistical* I.5.4 [**Pattern Recognition**]: Applications – *text processing*

General Terms

Algorithms, Experimentation

Keywords

Web-based Assessment, Statistical Fact Analysis, Content Filtering.
.

1. INTRODUCTION

In recent years, the volume of content on the web has increased dramatically. This has been fueled by the growing ease to author and publish content on collaborative environments such as online discussion forums, weblogs and wikis. In turn, the diversity among internet users has grown, and this diversity has also reflected on the quality of the content being published online. This has increased the motivation for tools to help users consuming this rich content, and make better use of the knowledge posted online.

Collaborative intelligence has been one of the major tools for using the wisdom of the crowd in gauging the value and support for content. Collaborative filtering has been heavily used in rating online content [8, 11]. Several platforms for user generated content allow its users to boost or demote the confidence in content. In addition, several services have emerged to help users identify the most important content being published online. Such services allow users to rank content based on popularity. However the sheer volume of content being produced limits the effectiveness of collaborative filtering, rendering algorithms to automatically assess content more valuable.

In this work we present an algorithm to assess textual documents against the web content using a statistical approach. The algorithm is applied in the context of online textual documents. The algorithm extracts basic elements of the document and accumulates support for each fact using web search. The individual support for each fact is accumulated to evaluate a final support score for a given document.

The remainder of this paper is organized as follows. Related approaches in the literature are introduced in section 2. Section 3 details the statistical approach suggested, and section 4 outlines the evaluation of this approach. Conclusions and some considerations for future research are outlined in section 5.

2. ASSESSMENT OF TEXTUAL CONTENT

Automatic assessment of textual content has attracted recent interest due to dramatic growth of online content. Several aspects are studied in this context including authors' interactivity trends in collaborative content [15], ranking of weblogs based on credibility scores [4], automatic scoring of online discussion forums content [17, 18] and surfactant features as quality indicators [1].

Viegas et al. [15] have studied cooperation and conflicts between Wikipedia authors by visualizing history flows. They have shown that their tool successfully monitored the evolving behavior of the Wikipedia community. Juffinger et al. [4] developed a weblog credibility ranking function by measuring the similarity against a previously prepared and verified news corpus. Despite reporting an 83% average precision on 14 blogs, a more exhaustive evaluation is required to address the potential generalization capabilities for their approach. Weimer et al. [18] proposed an algorithm for automatic assessment of online discussion post quality. They suggested a set of features that ranged from surface features, such as Capital Word Frequency, to more linguistically intense features, such as lexical and syntactic features. A trained

SMUC'10, October 30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0386-6/10/10…$10.00.

SVM was used to binary classify posts into either *good* or *bad* ones. Wanas et al.[17] have also studied scoring online forum posts using language-independent features extracted from the textual content of the post. They used embedded web links, punctuation occurrence patterns, relative post length, relative posting time, overlap between posts, emotion icons and other text-analysis based observations to combine different twenty two features used to score the post. These two approaches assume a set of dependent short documents, such as forum posts, which is not generalizable to different forms of online content. Gelman and Barletta [1] combined spelling errors and online search hit count to cook a quick quality indicator for web sites. Initial results have shown a correlation between spelling errors and content quality. However, the accuracy in more quality user content has not been discussed.

Several approaches have assessed the quality of Wikipedia content from a variety of different perspectives. Stvilia et al. [13] proposed seven metrics for assessing Wikipedia articles quality combined from eighteen different features. Some of these features can be directly extracted from a Wikipedia article including the number of unique editors, article length, and the number of reverts. Other features are derived from the whole dataset like connectivity. Their metrics distinguish featured Wikipedia articles with a higher value compared to random articles. Zeng et al. [19] used revision history of the article to estimate a trustworthiness measure. Under many simplifying assumptions, their statistical approach has shown promising results to model article's trust in terms of history information. Hu et al. [3] proposed three models to assess Wikipedia article's quality. The basic model depends on the principle that the quality of a document is directly proportional to the authority of contributing authors. That is to say that the overall quality of an article is the aggregation to all the contributing authors' authorities. Author's authority, in turn, is measured as an aggregation of his/her articles' quality. An iterative implementation is used to implement this approach. The second model is based on the PeerReview of articles. In this model, they consider the authority of article's editors. These editors are assumed to review a certain version of the article before modifying it, and this reflects on the final quality of a given article. However, not all articles undergo peer review, and in such case a probabilistic model to select reviewers is suggested in the third model, ProbReview. Similarly Lim et al.[6] consider each author's contribution into account. Blumenstock [2] reported his observations about the relation between word count and content quality in Wikipedia articles. He reported an accuracy of 96.31% with 2000 words threshold to binary classify Wikipedia articles as either *random* or *featured.* In turn, featured articles are those with more words. He also used more sophisticated classification models and presented the results. Despite the simplicity of the measures suggested, it outperforms other approaches [19, 12]. Other approaches modeled quality of Wikipedia articles in terms of accessibility-related information, citation-based analysis and the likes [7, 5]. However, common to all such approaches is the dependence on structural information, such as the one existing in Wikipedia which doesn't generalize to general online documents.

In the context of general web documents, Joshua et al. [9] have addressed the issues related automatic identification and measuring of the cohesiveness features of a web page. Through identifying the topics in a web page, they have developed a metric to calculate the cohesion among these topics. The more related the topics are, the higher is the cohesiveness value of the web page and, thus, the higher the quality. Using fully language-independent techniques in this approach make it sensitive to noisy content of a web page. In this paper, we aim to extend this work to assess the value of online documents

3. STATISTICAL WEB-BASED CHECKING OF TEXTUAL DOCUMENTS

A document is composed of a sequence of sentences, which in turn are composed of a sequence of phrases. Among these phrases are factual phrases that start and end with a noun (noun-to-noun phrase). As the proposed approach is a web-search based approach, and as web search usually is conducted using nouns, nouns are considered anchor points that are able to identify the context of the document; hence noun-to-noun phrases are considered potential facts. While a document can be represented with many facts, only a limited number are considered core elements. It is assumed that the web support of these core facts governs the overall support of the document. Based on this assumption, the work in this paper presents an algorithm to assess the web support of a textual document. The basic building blocks of this algorithm are (i) fact selection, (ii) individual fact assessment, and (iii) document score aggregation. In the following we shed light on these different elements in detail.

3.1. Fact Selection

Extraction of noun-to-noun facts from a document is based on parsing the document using a Part-Of-Speech (POS) tagger [14]. That is to say that for a document *D*, Noun-to-Noun phrases are identified and represented as a the set $f_k(D)$ where *k* is an index for these facts. A single fact, $f_k(D)$ extracted from sentence s_i, can be represented as follows

$$f_k(D) = (n_{1k}, t_k, n_{2k}) \tag{1}$$

Where, n_{1k} and n_{2k} are the head and tail nouns of the fact f_k respectively, and t_k is a string connecting these two nouns. The set $f_k(D)$ $\forall k$ represents all the facts in document *D*. These facts are categorized into two classes based on the connection string t_k, (i) standard semantic relations and (ii) other non-standard relationships. Standard semantic relations are the set of noun-noun relationships used in developing WordNet. This is a set of seven relationships, namely (i) *Is-a,* (ii) *Kind-of,* (iii) *Superordinate-of,* (iv) *Has-a,* (v) *Part-of,* (vi) *Opposite,* and (vii*) The-same-as*. The significance of these relationships is that they represent core factual elements within a document. Template matching is used to identify the category of each fact f_k. These templates mainly depend on matching the distinguishing keywords of the relationship. On the other hand, non-standard facts are all other facts that do not match any of these templates. While this approach might not extract all possible facts in the document, they will produce enough candidates to identify the degree of web-support present for the document.

While extracting the set of representative facts we take into consideration that all sentences in the document should be part of the document representation. Hence the set of representative facts is composed of a fact from each sentence. On the other hand, sentences appearing earlier in the documents are usually more important to the context of document representation. In turn, facts representing sentences appearing earlier in the document are given a higher weight.

Six different weighting schemes are suggested to rank facts, they are:

$$w(f_k(D)) = \begin{cases} |t_k| / tf(n_{1k}) + tf(n_{2k}) \\ |t_k| / max\big(tf(n_{1k}), tf(n_{2k})\big) \\ k * |t_k| / max\big(tf(n_{1k}), tf(n_{2k})\big) \\ k \\ 1 / tf(n_{1k}) + tf(n_{2k}) \\ 1 / max\big(tf(n_{1k}), tf(n_{2k})\big) \end{cases} \quad (2)$$

Facts are ordered based on the weights, while taking into consideration that facts with standard semantic relations take priority. Selection is performed either globally or on a per sentence basis. A threshold, *Thres*, on the number of facts is used to select the representative facts of a given document *F(D)*.

3.2. Individual Fact Assessment

For each fact $f_k(D)$ extracted from the document *D*, a set of four queries are generated. These queries are (i) n_{1j}, (ii) n_{2j}, (iii) n_{1j} *and* n_{2j}, and (iv) n_{1j} *and* t_j *and* n_{2j}. The set of search queries are issued to the Bing search engine (http://www.bing.com), and the top URLs resulting from the search are selected. URLs are assigned weights $W(u_l,f_k)$ based on two factors (i) their occurrence frequency in the list (ii) URL rank by the search engine. The top ten URLs, based on the assigned weights, are crawled, and facts are extracted from their content.While using more URLs might potential improve the accuracy, it will increase the computational time required dramatically. The fact $f_k(D)$ is matched with the facts extracted from the URL u_l, and a support value is assigned as

$$support\big(f_k(D)\big) = \frac{\sum_l W(u_l,f_k) * M(u_l,f_k)}{\sum_l W(u_l,f_k)} \quad (3)$$

where

$$M(u_l, f_k) = \begin{cases} LCS_{Match(f_k(D))} & \text{if } f_k(D) \text{ matched in } u_l \\ -LCS_{Match(f_k(D))} & \text{if } \overline{f_k(D)} \text{ matched in } u_l \\ 0 & \text{otherwise} \end{cases}$$

And $LCS_{Match(f_k(D))}$ is the longest common ordered, but not necessarily successive, words sequence between the two facts, in order taking into account negations.

Negations are considered based on (i) explicit negation words and (ii) matching the whole relation and the lead noun only (n_{1k} and t_k). The later scenario assumes that when the same relation links the lead noun, n_{1k}, to two distinct nouns it reflects a negative match. For example, if the two facts *Bill was born in Seattle* and *Bill was born in Maryland* exists, the later fact demotes the support of the first fact. This scenario fails to consider if the second fact is *Bill was born in USA,* so a potential enhancement is to discover the relation between second nouns instead of general assumption that they are different.

3.3. Document score aggregation

The support could be generalized from a single fact to a sentence, s_i, composed of multiple facts $F(s_i)$. In turn, the support for the sentence, $support_s(s_i)$ can be defined as the weighted sum of the individual facts as follows

$$support_s(s_i) = \frac{\sum_{f_k(D)\in s_i} w\big(f_k(D)\big) * support\big(f_k(D)\big)}{\sum_{f_k(D)\in s_i} w\big(f_k(D)\big)} \quad (4)$$

where $w(f_k(D))$ is the weight assigned to every fact.

This can be further generalized to aggregate a score for a document *D* composed of sentences $S = \{s_i \forall i\}$ as the weighted summation of individual sentence support as follows

$$support_D(D) = \frac{\sum_i W_s(s_i) * support_s(s_i)}{\sum_i W_s(s_i)} \quad (5)$$

where $W_s(s_i)$ is the weight assigned to every sentence based on its position in the document.

4. EVALUATION

The evaluation of the approach is divided into three different experiments (i) evaluating the quality of fact extraction, (ii) evaluation of the assessment of document support, and (iii) evaluation of the performance of the algorithm. In the following we will each of these experiments in some detail.

4.1. Fact Extraction Evaluation

The quality of the automatic assessment approach suggested is highly related to the quality and quantity of the facts selected to represent a given document. To investigate the fact selection we use the *The New York Times Annotated Corpus* dataset [10]. This corpus contains over 1.8 million articles written and published by the New York Time, with over 650,000 article summaries written by library scientists. It is assumed that the facts present within the summary represent the most important facts in the document. In turn the level of recall of the facts within the summary is an indication of the performance of the fact weighting scheme used. We reorder the documents' facts with the six different weights suggested in section 3.1. These weighting schemes could be applied either globally or sentence-based. Applying the weights globally means that all the facts are ordered irrespective of their location in the document, while the sentence-based approach applies the order within each sentence, and the top candidates from each sentence are then ordered globally.

The performance of the recall using different threshold levels and different weighting schemes applied globally or sentence-based is illustrated in Figures 1 and 2 respectively. The diversity in performance using the different weighting schemes is significant when reordering globally, while it is limited when using a sentence-based approach. This is due to the limited number of facts in the later approach. It is worth mentioning that incorporating the overall rank of the fact in the weighting scheme improves the overall performance when considering global ordering. This is not matched with the sentence-based approach, where the diversity amongst the different approaches is limited.

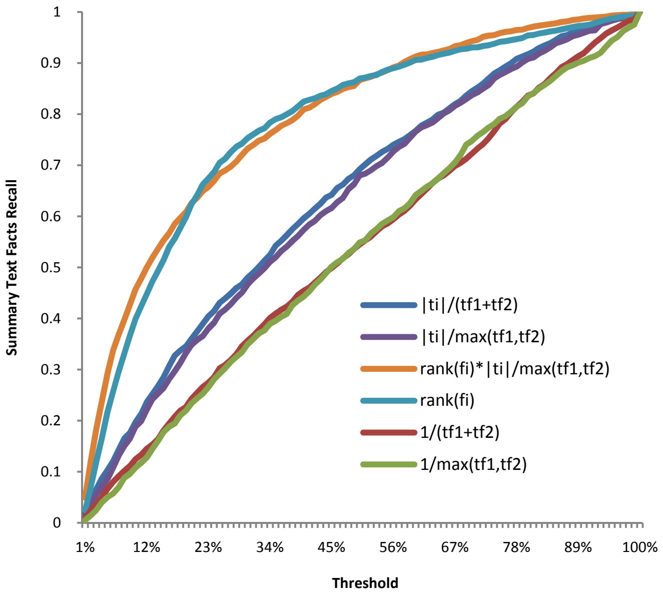

Figure 1. Recall of document summary facts using global ordering

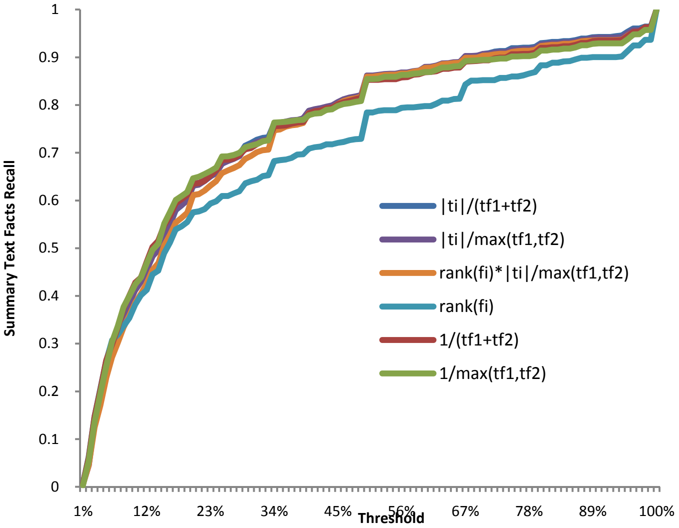

Figure 2: Recall of document summary facts using sentence based ordering

The performance outlines suggests that we can use the sentence-based approach due to its computational efficiency without loss of accuracy. In addition, selecting the threshold *Thres* to be be 30-60% allows the fact selection to cover 70-90% of the facts in the summaries.

4.2. Document Assessment Evaluation

To evaluate the performance of the online document assessment approach suggested we use a collection of Wikipedia articles. We crawled 100 Wikipedia articles, with 100 or more edits. The 100 articles included roughly equal number of featured articles, disputed articles, and randomly selected articles. Each roll back of edits is considered a new article.

For each sentence s_i, in an article, a set of representing facts $F(s_i) \in f(s_i)$ is selected, where $f(s_i)$ is a set of all facts in the sentence. Different values of the threshold, *Thres,* are used (30% - 60%) to limit the maximum number of facts selected. Moreover, the weight, $W_s(s_i)$, for the sentence, s_i, is selected to take equal values across the different sentences, or weighted based on the number of facts in each sentence. Average score is considered for all articles for the same revision number. The following subsections demonstrate results and conclusions of the experiments.

Thres percent of the extracted facts are used. The set of all extracted facts is the union of representative sets of all sentences. For each sentence s_i,

The average performance of featured, random and disputed articles for *Thres* = 30% of the number of facts generated for each document and the ranking model are illustrated in Figures 3 – 5. The figures outline the change in the score as edits are applied to the document. It can be noted that generally, the support score for the featured documents increases with the addition of new edits. In addition, the disputed articles tend to finally reduce in the overall score for support as edits are inserted in the document. Moreover, random articles in general tend to decrease in the quality with the edits inserted, yet not as significantly as the performance of the disputed articles.

The performance using the global or sentence-based ranking is marginally different, this was true for different values of threshold *Thres*. In turn, the selection of one fact per sentence will reduce the computational and storage requirements of the algorithm and hence is favorable. Sentences maybe weighted differently in the aggregation of the final document score, either uniformly or based on the number of facts extracted from each sentence. The results illustrate that the performance is not affected with this change.

The threshold, *Thres*, on the number of facts selected to represent the document has a significant impact on the performance as outlined in Figures 3, 6 and 7. The increased number of facts increases the distinction between the different article types. It is worth mentioning that the number of revisions required before the different article types are properly ordered relative to each other doesn’t change significantly with any of these changes in the implementation. The only exception is the use of global ranking at a threshold of 30%, where the number of revisions changes from around 40 to almost 70.

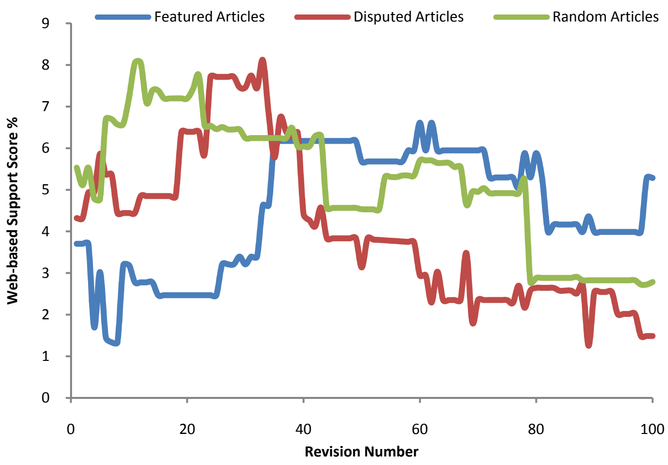

Figure 3: Performance of (a) Featured, (b) Random, and (c) Disputed articles with 100 revisions using *Thres = 30%* of sentences, $|F(s_i)| = 1$, $W_s(s_i) = 1$

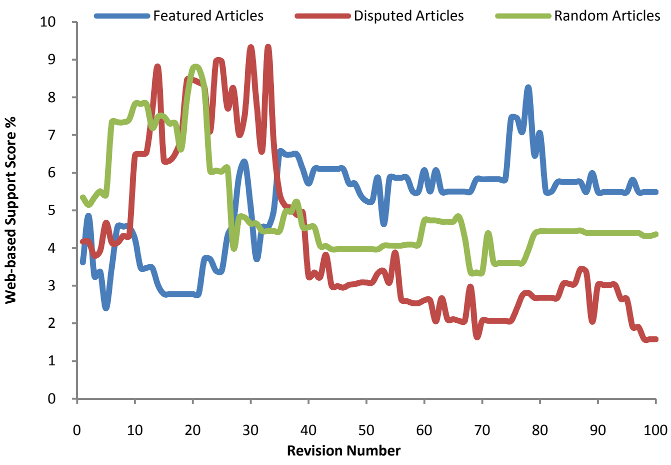

Figure 6: Performance of (a) Featured, (b) Random, and (c) Disputed articles with 100 revisions using *Thres = 40%* of sentences, $|F(s_i)| = 1$, $W_s(s_i) = 1$

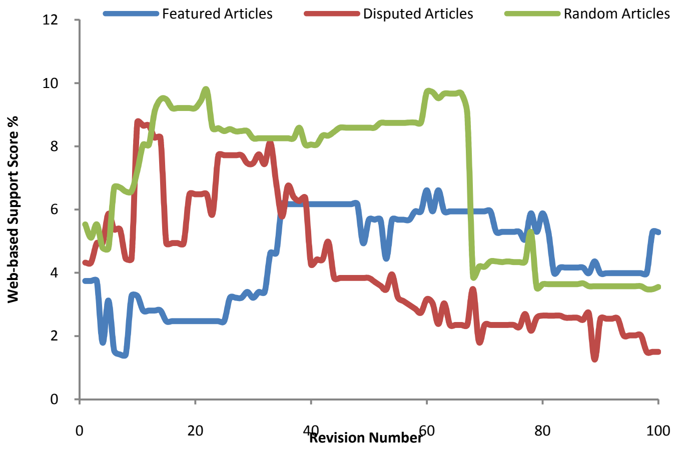

Figure 4:Performance of (a) Featured, (b) Random, and (c) Disputed articles with 100 revisions using *Thres = 30%* of sentences, $|F(s_i)| = |f(s_i)|$, $W_s(s_i) = 1$

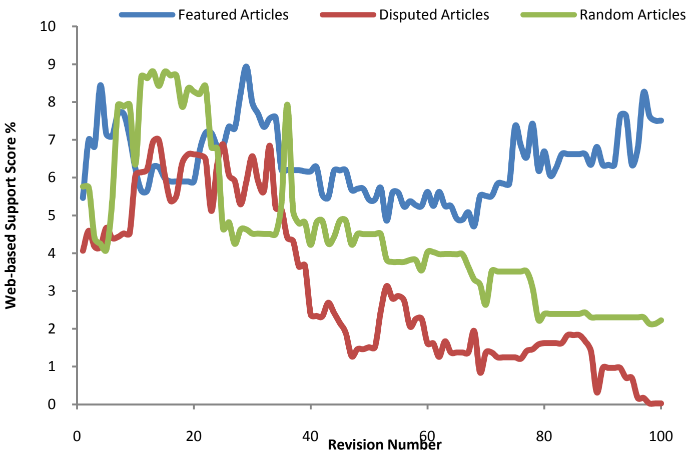

Figure 7: Performance of (a) Featured, (b) Random, and (c) Disputed articles with 100 revisions using *Thres = 60%* of sentences, $|F(s_i)| = 1$, $W_s(s_i) = 1$

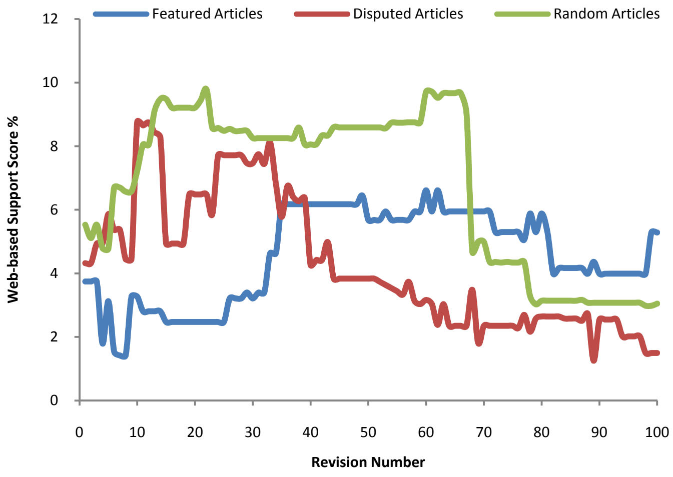

Figure 5:Performance of (a) Featured, (b) Random, and (c) Disputed articles with 100 revisions using *Thres = 30%* of sentences, $|F(s_i)| = |f(s_i)|$, $W_s(s_i) = |F(s_i)|$

4.3. Performance

In order to evaluate this approach a dataset is constructed to cover both true and false statements. News data is the best approximation of true statements since it is a factual by nature. 5529 documents of New York Times data extracted from LDC catalog LDC2008T25 [16] are selected to represent true facts. False facts, on the other hand, are collected from online true/false general knowledge quizzes. *Dalai Lama live in China*, *Bill Gates was born in Iran*, and *Egypt lies in Europe* are examples for false facts experimented.

Figure 8 shows a histogram of 843 documents resulting in a non-zero score. Only two documents that have very short text got a unit score. The distribution of the scores of factual documents is different from false documents. The score for false documents is generally more skewed towards low vales compared to factual news documents. The mid-range of scores (0.3-0.6) are assigned to factual documents. However, the other ranges of scores illustrate an ambiguity in clearly distinguishing the quality of the document. It is worth noting that some of the false facts were evaluated with a positive aggregate score since these type of questions always contain a true segment and only a limited false component. In most cases, the module extracts facts of the right part so they are matched versus the web content.

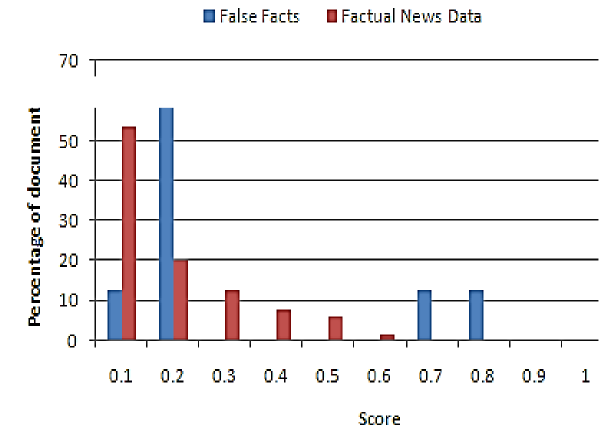

Figure 8: Distribution of non-zero score documents on New York Times data and false facts

The majority of the documents were assigned a score of zero, reflecting the inability of the proposed approach to assess these documents. Among these documents, 1307 return URLs from the search queries issued, but these URLs lack matching facts. The scores have shown above shows different score ranges for the same type of documents that assumed to be factual. However, all the score didn't exceed 0.7 score which is a controlled score range.

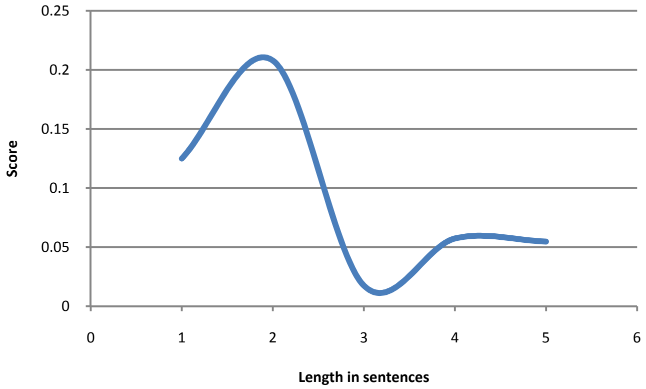

Figure 9: Average performance of fact checking module with varying the document length

It is noted that the length of document is a factor that affects the score range. In order to assess this aspect, a collection of 200 documents collected from the first section of Wikipedia English articles is used. Different documents are extracted from the same original Wikipedia article with different lengths in sentences. An experiment is run over this data set and average scores of each length are plotted as shown in Figure 9. On the average, the score decreases with increasing the length of the document. This behavior is expected since with longer documents more facts are extracted and in turn more URLs are obtained. In turn, the probability of matching facts extracted from the web search results become lower and the overall score is demoted.

5. CONCLUSION AND DISCUSSION

In this paper we presented a statistical approach to assess textual documents using a web-based fact checking. Key facts, that represent basic elements of the document, are extracted. Thereon, search is used to validate the web support of these elements. Then, an overall score is aggregated from the individual support for each element to indicate the document web support. An experimental study is performed to investigate the effect of such parameters using a collection of factual news articles and false facts dataset. Evaluation articles are selected to represent documents with different levels of quality to illustrate the behavior of the algorithm with varying levels of web content quality. The increase in the length affects the outcome of this algorithm.

Different parameters affect fact extraction process, mostly related to thresholds for selection and ranking of facts. An experimental study is performed to investigate the effect of such parameters using a collection of Wikipedia articles. Evaluation articles are picked up of different qualities to show the behavior of the algorithm with varying levels of web content quality.

The algorithm has illustrated an ability to distinguish between different articles based on quality. The increase in the number of facts selected to represent a document hugely affects the outcome of this algorithm. Moreover, while quality of the different articles could start from similar levels, it quickly diverges to distinctive levels with the increase in the number of edits.

While the proposed approach seems promising, many improvements can be made. Anchor points identification within the document can be enhanced by extending selection beyond nouns. Combinations from nouns and verbs should play a better role in both identifying anchor points and discovery of relationships that tie these anchor points. In addition, literature of open relationship extraction is expected to add a value to the template matching approach currently in use to discover relationships between nouns.

6. ACKNOWLEDGMENTS

This research is conducted during an internship granted to the first author in Cairo Microsoft Innovation Lab.

7. REFERENCES

[1] Gelman, I.and Barletta, A. 2008.“A quick and dirty" website data quality indicator. *In Proceedingsof the 2nd ACM workshop on Information Credibility on the Web* (Napa Valley, CA, USA October 30, 2008) WICOW08.ACM New York, NY, USA, 43-46.

[2] Blumenstock, J. 2008.Size matters: word count as a measure of quality on Wikipedia *,In Proceedings of the 17th International Conference on World Wide Web* (Beijing, China, April 21-25, 2008) WWW2008. ACM, New York, NY,USA, 1095-1096.

[3] Hu, M., Lim, E., Sun, A.,Lauw, H., and Vuong, B. 2007.Measuring article quality in Wikipedia: models and evaluation, *In Proceedings of the 16th ACM Conference on*

Information and Knowledge management (Lisbon, Portugal, November 6-9, 2007) CIKM 2007.ACM, New York, NY, USA,243-252.

[4] Juffinger, A., Granitzer, M., and Lex, E. 2009.Blogcredibility ranking by exploiting verified content, *In Proceedings of the 3rd workshop on Information Credibility on the Web* (Madrid, Spain, April 20, 2009) WICOW09.ACM, New York, NY, USA, 51-58.

[5] Lih, A. 2004. Wikipedia as Participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. *In Proceedings of Fifth International Symposium on Online Journalism* (Austin, TX, USA, April 16-17, 2004)16-17.

[6] Lim, E., Vuong, B., Lauw, H., and Sun, A. 2006.Measuringqualities of articles contributed by online communities. *In Proceedings of the 2006IEEE/WIC/ACM International Conference on Web Intelligence* (Hong Kong, China, December 18-22, 2006) ACM New York, NY, USA, 81-87.

[7] Lopes, R., and Carricco, L. 2008.On the credibility of Wikipedia: an accessibility perspective *In Proceedings of the 2nd ACM workshop on Information credibility on the web* (Napa Valley, CA, USA October 30, 2008) WICOW08.ACM New York, NY, USA, 27-34.

[8] Papagelis, M., Rousidis, I., Plexousakis, D., and Theoharopoulos, E. 2005. Incremental collaborative filtering for highly-scalable recommendation algorithms, *In Proceedings of the International Symposium on Methodologies of Intelligent Systems(*Saratoga Springs, NY, USA, May 25-28 2005)ISMIS'05, Springer Verlag, New York, NY, USA 553-561.

[9] Pun, J., and Lochovsky, F. 2005. Finding high-quality web pages using cohesiveness, *In Proceedings of the 2005 International Conference on Information Quality* (Houston, TX, USA, September 19-23, 2005) IQ2005. MIT Press, Cambridge, MA, USA.

[10] Sandhaus, E, 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.

[11] Sarwar, B.,Karypis, G., Konstan, J., and Reidl, J. 2001. Item-based collaborative Filtering recommendation algorithms", *In Proceedings of the 10^th^International Conference on World Wide Web* (Hong Kong, China, May 1-5, 2001) WWW2001. ACM, New York, NY, USA 285-295.

[12] Stvilia, B., Twidale, M., Gasser, L., and Smith, L. 2005.Information quality discussions in Wikipedia, *In Proceedings of the 2005 International Conference on Knowledge Management*(Charlotte, NC, USA, October 27-28, 2005) ICKM2005.101-113.

[13] Stvilia, B., Twidale, M., Smith, L., and Gasser, L.,2005. Assessing information quality of a community-based encyclopedia, *In Proceedings of the 2005 International Conference on Information Quality*(Houston, TX, USA, September 19-23, 2005) IQ2005. MIT Press, Cambridge, MA, USA, 442-454.

[14] Toutanova, K., Klein, D., Manning, C., and Singer, Y., 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *In Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (Edmonton, AB, Canada May 27 – June 1, 2003)HLT-NAACL 2003. ACL, Morristown, NJ, USA, 252-259.

[15] Viegas, F., Wattenberg, M., and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations, *In Proceedings of the SIGCHI Conference on Human factors in computing systems* (Vienna, Austria, April 24-29, 2004) CHI 2004. ACM, New York, NY, USA, 575-582.

[16] Vorhees, E., and Graff, D., 2008. AQUAINT-2 Information-Retrieval Text Research Collection. Linguistic Data Consortium, Philadelphia.

[17] Wanas, N., El-Saban, M., Ashour, H., and Ammar, W., 2008.Automatic scoring of online discussion posts, *In Proceedings of the 2nd ACM workshop on Information Credibility on the Web* (Napa Valley, CA, USA October 30, 2008) WICOW08.ACM New York, NY, USA, 19-26.

[18] Weimer, M., Gurevych, I., and Muhlhauser, M., 2007. Automatically assessing the post quality in online discussions on software, *In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*(Prague, Czech Republic, June 23-30, 2007) ACL2007.ACL, Morristown, NJ, USA, 2007, pp. 125-128.

[19] Zeng, H., Alhossaini, M., Ding, L., Fikes, R., and Mcguinness, D., 2006. Computing trust from revision history, *In Proceedings of the 2006 International Conference on Privacy, Security and Trust* (Markam, Ontario, Canada, October 30 – November 1, 2006) PST 2006. ACM, New York, NY, USA, 8.