Appraisal PageRank for Arabic and English Blogs: Comparative Study

Mustafa M. Noaman
Computer Science Department, IT & CS Faculty,
Yarmouk University
Irbid, Jordan

Belal M. Abuata
Computer Information Systems Department, IT & CS Faculty,
Yarmouk University
Irbid, Jordan

ABSTRACT
A comparative study is made between Arabic and English Blogs. The difference between Arabic and English blogs starts with their Pagerank. The question to answer in this paper is "Do the Arab blogger benefits from the search engine optimizer properties or not?" In other words does the user cares for Pagerank? This question rises after the fame that the social media in general and blogs in specific have played during the political changes in Middle East region. Using a corpus of Arabic & English blogs, their content & link properties that effect the Pagerank score to answer the question (to prove or deny whether our assumption is true or not). The results obtained showed that there is a huge difference between Arabic & English blogs in terms of Page rank and content.

General Terms: Web Blogs.

Keywords: *Pagerank, Arabic Blogs, English Blogs, Arab Bloggers, Search Engine Optimizer*.

1. INTRODUCTION
Business Week magazine stated on its May/2005 cover "Blogs will change your business". Blog or Web Blog is a type of web site used for publishing purposes in the internet. It is considered as a part of the collaborative environment which emerged in the last 10 years due to rich & ease of use of hypertext to connect writers & readers, who interested in certain topic rather than using other tools introduced by internet ex. (chat, e-mail, discussion forums, etc.).This is due to the simplicity, permanent linking to specified topic and opportunity to express opinions about a certain topic and diversity of topics covered by Blogs [1, 2, 3]. In February 2011 there were over 156 million public blogs in the World Wide Web [4]. Arab internet users are found mainly in the Middle East region and 5% of world population and 4% of the world internet users. Arabic web materials do not exceed 2% where most of these materials are blogs [1]. Arab bloggers numbers are growing to be more effective as the use of the internet continue to grow in the Middle East region. The majority of Arabic blogs are abandoned, inactive or may contain ordinary content. Several studies estimated the number of significant blogs in thousands [1, 5]. Pagerank (PR) was proposed & developed by Google's founders (Larry Page and Sergey Brias) as part of a research project about a new kind of search engine. The idea was to order information on the web in a hierarchy by link popularity [6]. Pagerank is defined as a numeric value that represents how important a page is on the web or it's Google's method for measuring a page's importance. Pagerank is Google's way of deciding a page's importance. Pagerank is one of the factors that determine a page's ranking in the search results [6]. The next section will be devoted to talk about the emergence of Pagerank, English & Arabic blogs, the relation between blogs & Google Pagerank and the characteristics that affect it. The 3rd section will describe our Corpus and Pager Rank features in details.

2. LITERATURE REVIEW
Pagerank can be seen as a model of user behaviour. It assumes that there is a random web surfer; he/she starts from a web page randomly. Web surfers usually keep clicking on the "forward" links, not "back" but when the time passes they get bored and chose another random web page. Pagerank computes the probability that the random web surfer visits a page.
The possibility of a web page being visited is determined by several factors:

- The importance of the webpage. The most important factor that increases the possibility of a web page being visited by web surfer.
- The total number of web pages that point (inbound links) to the web page.
- The importance and the forward (outbound) link number of each of these web pages. [7]

The original Pagerank algorithm was proposed by Google's founders Lawrence Page and Sergey Brin.
It is given by:

$$PR(A) = (1-d) + d(PR(Ti)/C(Ti) + ...PR(Tn)/(Tn)) \quad (1)$$

Where:
PR(A) is the Pagerank of page A.
PR(Ti) is the Pagerank of pages Ti which link to page A, where: $1 \leq i \leq n$.
C(Ti) is the number of outbound links on page Ti.
d is a damping factor which can be set between 0 and 1. [7]
The sum of the weighted Pageranks of all pages Ti is multiplied with a damping factor d with a value between 0 and 1. Hence, the part of Pagerank benefit for a page by another page linking to it is reduced [7]. One drawback of Pagerank equation is that it doesn't rank web sites as a group but it's calculated for each page individually. Also the Pagerank of page A is defined by the Pagerank of pages which link it (inbound links) [8]. The Pagerank of pages Ti that link to page A does not affect the Pagerank of page A in regular way. The Pagerank of a page T is measured by the number of outbound links C (T) on page T [8]. A conclusion for equation (1) shows that when the number of outbound links page T has the C(T) increases, the less will page A benefit from a link to it on page T. In contrast, when the weighted Pagerank of pages Ti increases, this will result in an additional inbound link for page A that will always increase page A's PageRank [7].
The new Pagerank algorithm presented as follows:

(2)

$$r(p) = \alpha \times \sum_{q:(q,p)\in\varepsilon} r(q)/w(q) + (1-\alpha)\times 1/N$$

Where:

r(p) is the Pagerank value for a web page p.
w(q) is the number of forward links on the page q.
N is the total number of web pages in the Web.
α is the damping factor. [8]

Therefore, the high Pagerank for a specific page occurs in two situations:

- If there are many pages that point to it.
- If there are some pages with high Pagerank pointing to it.

But it is vulnerable to link-based spamming techniques.

Prior to blogs there were many forms for communicating internet users interested in certain topics such as e-mail lists, bulletin board systems and forums. Blogging evolved from the concept of online diaries where people write about their daily activities and stories. Justin Hall is considered one of the pioneer bloggers who documented his life using blogs. He stated from 1994 and kept blogging for 10 years [9, 10, 11]. Citizens of the Arab world wanted to show their opinions and journalism skills over the internet that provide less control and wider audience. Arab bloggers like other bloggers around the world used the blogs for publishing raw & uncensored materials over the internet and they are not necessarily journalists or following guidelines, press laws or ethical codes [5, 11]. The number of Arabic bloggers increased in recent years due to the political changes and wars in Middle East region which embraces the need for showing opinions and discussing political topics of interest to audience [5]. Many websites offered the power of blogging for free like Maktoobblog.com and these websites encourage Arabic people to express their opinions in many topics like politics, religion, news, arts/literature, women's issues/rights, minority issues/rights, pop culture (music, TV, movies), sports, technology etc.[12]

3. Methodology

This section describes the blogs corpus used and Pagerank features in details.

3.1 Blogs Corpus

The blogs corpus contains around 3000 Arabic and English blogs in different areas such as (sport, art, political, religion). The blogs state were one of three states: active or popular or abandoned, through the period of June to September 2011, using Google, and Yahoo search engines (figure 1). Then the Pagerank features computed for every blog in the corpus, using different web pages analyzers such as A1 Website Analyzer1, A1 Keyword Analyzer2 and SeoQuake3. The Pagerank features are: URL Length, Title Length, META Description Length, and Link Popularity through important search engines such as: Internal/External Links Count, Words Count, and file size [14].

Blog-URL	In-Lin	Out-	Blog-	URL-	Title-	Blog-Worl
http://weddingdressessaletop620bestseller.blogspot.com/	6	9	1442	48	7	1344
http://cristaldelicacy.blogspot.com/	57	53	109	29	3	1375
http://top890bestsellingwineaccessories.blogspot.com/	6	8	87	46	7	1340
http://outdoor-storagecabinet.blogspot.com/	35	15	109	36	3	1461
http://lawsoftwares.blogspot.com/	58	49	63	26	2	1377
http://annetteelisanatalie.blogspot.com/	26	197	82	33	1	1422
http://blackfridayblackfridaygarmingpstop210.blogspot.com/	6	9	173	51	9	1665
http://camlux.blogspot.com/	40	10	112	20	2	909

Figure 1: Blogs corpus.

3.2 Pagerank Features

The following is brief description of the Pagerank features used:

1. URL Length: The number of characters in the URL of the blog [13], for example: *abc.maktoobblog.com* contains 15 characters.
2. Title's Length: The title a text appears in the beginning of the blog in the browser and in search engines when searching for the blog and measured in number of words. Example: when search for "google bloog" in Google search engine, the blog will be shown as "Official Google Blog" [13].
3. META Description Length: The Meta description is intended to be a brief and concise summary of the page content and can be used by search engines or directories and measured in number of words [13]. For example: when search for "gmail blog" in Google search engine, the blog will be shown with Meta description: "Official Google blog for the web-based mail service, with news, developments, and productivity tips."
4. Link Popularity: Link popularity means the number of links that the search engine points to specific web page or blog [14, 13], for example: The popularity of hemmatlashin.maktoobblog.com in Yahoo search engine is 1615 which means that 1615 links from Yahoo search engine point to the blog.

[1] http://www.microsystools.com/products/website-analyzer/
[2] http://www.microsystools.com/products/keyword-research/
[3] http://www.seoquake.com/

5. Internal/External Links Count: The internal (inbound) / external (outbound) link is considered as hyperlink that is a reference element in a page to another section of the same page or to another page that may be on or part of the same blog or domain of the internet. Links are considered either "external" or "internal" depending on perspective. A link to a page in the same domain is considered internal as shown in figure 2, where if outside the same domain is considered external as shown in figure 3 [14,13].

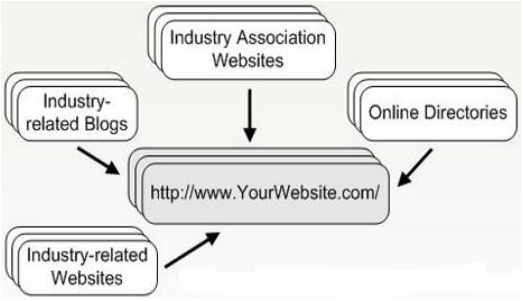

Figure 2: Internal links.

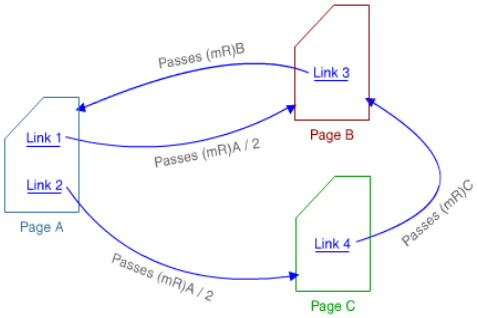

Figure 3: External links.

6. Words Count: the words count measures the number of words in the blog [13], for example: ayidh11.maktoobblog.com blog contains 14321 words.

7. Page Size: The size of the page' HTML and measured in KB (Kilo-Bytes) [13], for example: The page size for reachhumour.blogspot.com blog is 413.7 KB.

4. Results

The Decision tree J48 algorithm is used on our corpus using WEKA data mining tool, as a tester for our hypothesis about appraisal Pagerank features for Arabic blogs.

The analysis process of the Arabic and English blogs with their feature of Pagerank using WEKA tool and express figures by parallel coordination; figure 4 shows the relation between the language and number of characters in the URL, which shows that the English blogs use more characters in their URL than Arabic blogs.

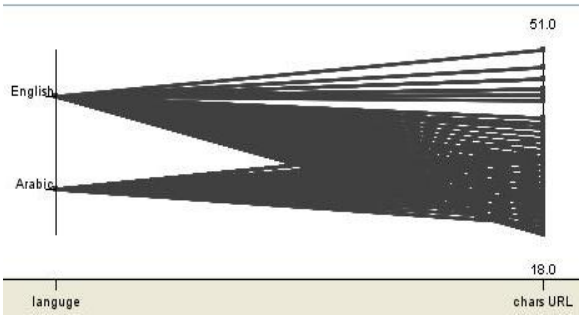

Figure 4: Parallel coordinates for URL and language used.

Figure 5 shows the relation between the language of the blogs and the length of title. This is due to the high weight caused by length of title on search engines and Pagerank considerations. The results showed that the English blogs longer title's lengths than Arabic blogs.

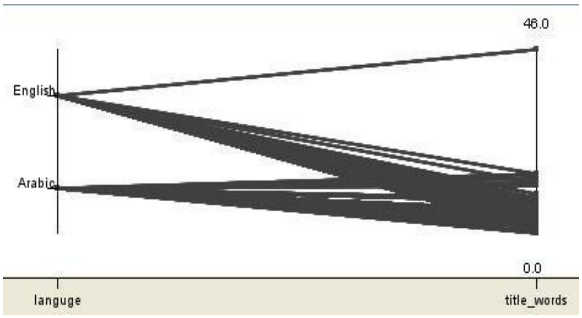

Figure 5: Parallel coordinates for title and language used.

Figure 6 shows the relation between the language of the blogs and the Meta description; as it provides a summary of the web pages in search engines considerations. The results showed that the English blogs use more words in their Meta those in Arabic blogs.

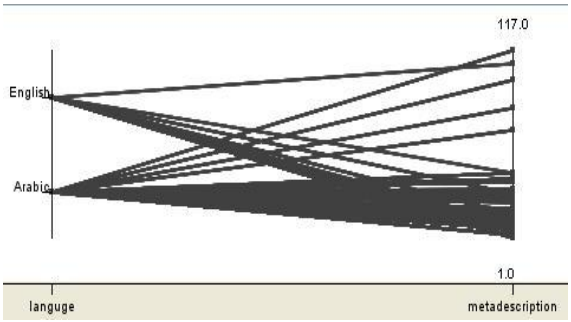

Figure 6: Parallel coordinates for Meta and language used.

Figure 7 shows the relation between the language of the blogs and the Link Popularity; which is considered a main reason in Yahoo and MSN to crawl the pages. It shows that the English blogs are more popular than Arabic blogs.

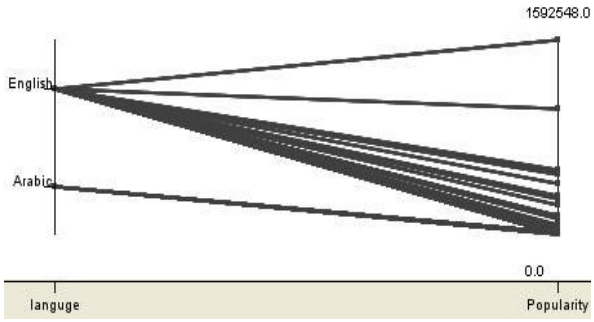

Figure 7: Parallel coordinates for Link Popularity and language used.

Figure 8 shows the relation between the language of the blogs and the Internal Link; which is considered as one of the main reasons in Google to crawl the pages. It shows that the English blogs have more internal links than Arabic blogs.

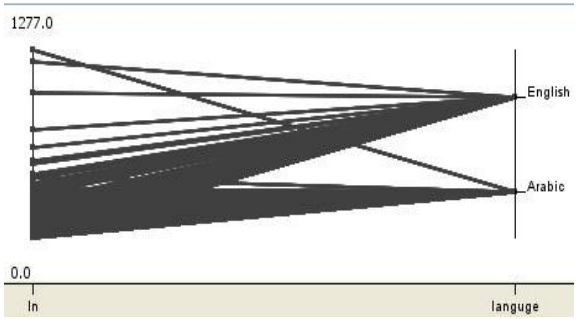

Figure 8: Parallel coordinates for Internal Link and language used.

Figure 9 shows the relation between the language of the blogs and the External Link; which is considered as one of the main reasons in Google to crawl the pages, which show that the English blogs have more external links than Arabic blogs.

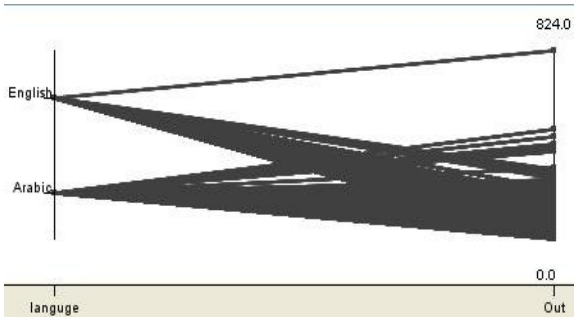

Figure 9: Parallel coordinates for External Link and language used.

Figure 10 shows the relation between the language of the blogs and the words count ; which is considered as a main feature affected in the Pagerank, and constitutes the highest score in Pagerank considerations. It shows that the English blogs have higher word counts compared to Arabic blogs.

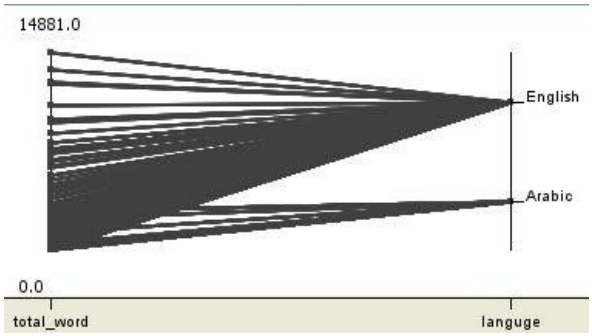

Figure 10: Parallel coordinates for Words count and language used.

Figure 11 shows the relation between the language of the blogs and the page size; which show that the English blogs have bigger page size than Arabic blogs.

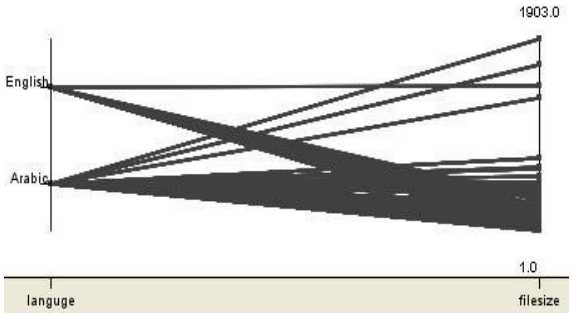

Figure 11: Parallel coordinates for page size and language used.

The accuracy score is 99% using Decision Tree J48. Hence, the conclusion acquired from tester that the values of Pagerank features is difference for Arabic and English blogs. The high accuracy indicate that owners of Arabic blogs do not benefit from search engines optimizer (SEO) advantages to improve both the quality and the Pagerank for Arabic blogs.

Table 1 show Precision, Recall, F- measure, True Positive, and False Positive, for our blogs corpus.

Table 1: The results of our blogs corpus.

	English blogs	**Arabic blogs**
True Positive	1	0.995
False Positive	0.005	0
Precision	0.995	1
Recall	0.997	0.997
F- Measure	0.997	0.997

5. Conclusions

The idea of appraising the Pagerank of Arabic and English blogs emerged from the increased attention that the social media got in the last years due to the political changes in the Middle East region as they provide the users the chance for expressing opinions and discussing issues in different areas. The results showed that there is a huge difference between Arabic & English blogs in terms of Pagerank and

content, as a result of Arabic bloggers don't utilize SEO guidelines to improve the Arabic blogs Pagerank which leads to more readers & attention to the blog. The future work will be to improve the Arabic blogs in terms of Pagerank & content and present the SEO guidelines that the Arabic blogger should follow to improve the current state of Arabic blogs.

6. References

[1] Boulos M., Maramba I., Wheeler S., "Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education," *BMC Medical Education*, vol. 6, no. 41, pp. 1-8, August 2006.

[2] Dwyer P., "Building Trust with Corporate Blogs," *ICWSM 07*, Colorado, USA, pp. 1-8, 2007.

[3] Godwin-Jones R., "EMERGING TECHNOLOGIES Blogs and Wikis: Environments for On-line Collaboration," *Language Learning & Technology*. vol.7, no. 2, pp. 12-16, May 2003.

[4] "BlogPulse," The Nielsen Company, www.blogpulse.com/ February 2011.

[5] Hamdy N., "Arab Citizen Journalism in Action: Challenging Mainstream Media, Authorities and Media Laws," *Westminster Papers in Communication and Culture*, vol. 6, no. 1, pp. 92-112, 2009.

[6] Langville A., and Meyer C., Google's PageRank and Beyond: The Science of Search Engine Rankings, *Princeton University Press*, ISBN: 0691122024, 2006.

[7] Sobek M., "A Survey of Google's PageRank," DOI= http://pr.efactory.de/ , 2002, (accessed May 2011).

[8] Hristidis V., "Random Walks in Ranking Query Results in Semistructured Databases," DOI=http://users.cis.fiu.edu/~vagelis/presentations/RandomWalks.ppt., 2003, (accessed May 2011).

[9] Gallagher D., "TECHNOLOGY; A Rift Among Bloggers," The New York Times, DOI= http://www.nytimes.com/2002/06/10/business/technology-a-rift-among-bloggers.html, June 2002, (accessed May 2011).

[10] Harmanci R., "Time to get a life - pioneer blogger Justin Hall bows out at 31," San Francisco Chronicl, DOI=http://www.sfgate.com/cgibin/article.cgi?file=/c/a/2005/02/20/MNGBKBEJO01.DTL, 2005, (accessed June 2011).

[11] Khoja S., "An RSS Feed Analysis Application and Corpus Builder," *Interface on the Internet*, DOI= http://www.elda.org/medar-conference/pdf/73.pdf, 2009, (accessed June 2011).

[12] Etling B., Kelly J., Faris R., and Palfrey J, "Mapping the Arabic Blogosphere: Politics, Culture, and Dissent," *Berkman Center Research Publication*, June 2009.

[13] "SEO tutorial: Introduction to SEO," FlamingoSoft website, DOI= http://seo-tutorial.seoadministrator.com/, December 2010, (accessed May 2011).

[10] Kerchove C., Ninove L., Dooren P., "Maximizing PageRank via outlinks," *Linear Algebra and its Applications*. vol. 429, pp. 1254–1276, 2008.