Computational Linkuistics: word triggers across hyperlinks

Dragomir R. Radev[1,2], **Hong Qi**[1], **Daniel Tam**[2], **Adam Winkel**[2]
[1]School of Information and [2]Department of EECS
University of Michigan
Ann Arbor, MI 48109-1092
{radev,hqi,dtam,winkela}@umich.edu

Abstract

It is known that context words tend to be self-triggers, that is, the probability of a content word to appear more than once in a document, given that it already appears once, is significantly higher than the probability of the first occurrence. We look at self-triggerability across hyperlinks on the Web. We show that the probability of a word w_j to appear in a Web document d_i depends on the presence of w_j in documents pointing to d_i. In Document Modeling, we will propose the use of a correction factor, R, which indicates how much more likely a word is to appear in a document given that another document containing the same word is linked to it.

1 Introduction

Given the size of the Web, it is intuitively very hard to find a given page of interest by just following links. Classic results have shown however, that the link structure of the Web is not random. Various models have been proposed including power law distributions (the "rich get richer" model), and lexical models. In this paper, we will investigate how the presence of a given word in a given Web document d_i affects the presence of the same word in documents *linked* to d_i. We will use the term *Computational Linkuistics* to describe the study of hyperlinks for Document Modeling and Information Retrieval purposes.

1.1 Link structure of the Web

Random graphs have been studied by Erdős and Rényi (Erdős and Rényi, 1960). In a random graph, edges are added sequentially with both vertices of a new edge chosen randomly.

The diameter d of the Web (that is, the average number of links from any given page to another) has been found to be a constant (approximately $18.59 = logN/logk$, where N is the number of documents on the Web and k is the average document *out-degree* (i.e., the number of pages linked from the document). This result was described in (Barabási and Albert, 1999) and is based on a corpus of 800 M web pages). This estimate of d would entail that in a random graph model, the size of the Web would be approximately 7^{19} which is 10 M times its actual size. Clearly, a random graph model is not an appropriate description of the Web. Instead, it has been shown that due to *preferential attachment* (Barabási and Albert, 1999), the out-degree distribution follows a power law. The preferential model makes it more likely that a new random edge will connect two vertices that already have a high degree. Specifically, the degree of pages is distributed according to $P(\Upsilon = k) \sim 1/k^{\alpha}$, where α is a constant strictly greater than 0. (Note this is different from $1/\alpha^k$, the distribution of out-degree on random graphs.) As a result, random walks on the Web graph soon reach well-connected nodes.

1.2 Lexical structure of the Web

Davison (Davison, 2000) discusses the *topical locality hypothesis*, namely that new edges are more likely to connect pages that are semantically related. In Davison's experiment, semantic and link distances between pairs of pages from a 100 K page corpus were computed. Davison describes results associating TF*IDF cosine similarity (Salton and McGill, 1983) and link hop distance. He reports that the cosine similarity between pages selected at random from his corpus is 0.02 whereas that number increases significantly for topologically related pages: 0.31 for pages from the same Web domain, 0.23 for linked pages, and 0.19 for sibling pages (pages pointed to by the same page).

Menczer (Menczer, 2001) introduces the *link-content conjecture* states that the semantic content of a web page can be inferred from the pages that point to it. Menczer uses a corpus of 373 K pages and employs a non-linear least squares fit to come up with a semantic model connecting cosine-based semantic similarity $\sigma(p_1, p_2)$ and the link distance $\delta(p_1, p_2)$ between two pages p_1 and p_2 (the shortest directed distance on the hy-

pertext graph from p_1 to p_2). Menczer reports that σ and δ are connected via a power law: $\sigma(\delta) \approx \sigma_\infty + (1 - \sigma_\infty)e^{(-\alpha_1 \cdot \delta^{\alpha_2})}$. σ_∞ represents noise level in similarity.

Menczer reports empirically determined values of the parameters of the fit as follows: $\alpha_1 = 1.8$, $\alpha_2 = 0.6$, and $\sigma_\infty = 0.03$.

Menczer's results further confirm Davison's observations that pages adjacent in hyperlink space to a given page are semantically connected.

Our idea has been to investigate the circumstances under which the semantic similarity between linked pages can be explained in terms of the presence of individual words across links.

1.3 Document modeling

In the computational linguistics and speech communities, the notion of a *language model* is used to describe a probability distribution over words. Since a cluster of documents contains a subset of an entire language, a document model is a special case of a language model. As such, it can be expressed as a conditional probability distribution indicating how likely a word is to appear in a document given some context (e.g., other similar documents, the topic of the document, etc.). Language models are used in speech recognition (Chen and Goodman, 1996), document indexing (Bookstein and Swanson, 1974; Croft and Harper, 1979) and information retrieval (Ponte and Croft, 1998).

Document models are a special class of language models. One property of document models is that they can be used to predict some lexical properties of textual documents, e.g., the frequency of a certain word. Mosteller and Wallace (Mosteller and Wallace, 1984) discovered that content words are "bursty" - the appearance of a content word significantly increases the probability that the word would appear again. Church and his colleagues (Church and Gale, 1995; Church, 2000) describe document models based on the distribution of the frequencies of individual words over large document collections. In (Church and Gale, 1995), Church and Gale compare document models based on the Poisson distribution, the 2-Poisson distribution (Bookstein and Swanson, 1974), as well as generic Poisson mixtures. A Poisson mixture is described by $P(x) = \int_0^\infty \phi(\theta)\pi(\theta, x)d\theta$, where $\pi(\theta, k) \approx \frac{e^{-\theta}\theta^k}{k!}$ for a given integer non-negative value of x.

Church and Gale empirically show that Poisson mixtures are a more accurate model for describing the distribution of words in documents within a corpus. They obtain the best fits with the Negative Binomial model and the K-mixture (both special cases of Poisson mixtures) (Church and Gale, 1995). In the Negative Binomial case, $\phi(\theta) = \frac{\theta^{N-1}e^{-\frac{\theta}{P}}}{P^N\Gamma(N)}$ (which is the Gamma distribution) whereas in the K-mixture, $\phi(\theta) = (1-\alpha)\delta(\theta) + \frac{\alpha}{\beta}e^{-\frac{\theta}{\beta}}$, where $\delta(x)$ is Dirac's delta function.

Our study focuses on modeling across hyperlinks. Documents linked across the web are often written by people with different backgrounds and language usage pattern.

1.4 Link-based document models

In Church et al.'s experiments, the documents being modeled do not have hyperlinks between them. When modeling hyperlinked corpora, it is important to decompose the document model into *link-free* and *link-dependent* components. The link-free component predicts the probability of a word w appearing in a document D regardless of the documents that point to D. The link-dependent part makes use of a particular incarnation of the link-content conjecture, namely *micro link-content dependency* (MLD), which we will propose in this paper.

1.5 Our framework

In traditional Information Retrieval, the main object that is represented, and searched, is the *document*. In our setup, we will be looking at the hyperlink between two documents as the main object to retrieve. If a page p_i points to page p_j via link l_k, then we will consider l_k as the object to index and the two pages that it links as features describing the link.

For our experiments, we used the 2-Gigabyte *wt2g* corpus (Hawking, 2002) which contains 247,491 Web documents connected with 3,118,248 links. These documents contain 948,036 unique words (after Porter-style stemming).

2 A link-based document model

It is well known that the distributions of words in text depend on many factors such as genre, topic, author, etc. Certain words with high content has been found to "trigger" other words to appear. Interestingly, the hyperlinks which connect the text on the Web may also affect the word distributions in the hypertext. For example, if page p_i that contains *education* points to page p_j, then we would expect a higher probability of seeing *education* in page p_j than in a random page. This experiment was designed to discover how the links between pages can trigger words and change the word distributions.

For each stemmed word in wt2g, we compute the following numbers:

PagesContainingWord = how many pages in the collection contain the word.

OutgoingLinks = the total number of outgoing links in all the pages that contain the word.

LinkedPagesContainingWord = how many of the linked pages contain the word.

For the latter two measures, only the links inside the collection were considered.

The probability of a word w appearing in a random page p_i is computed as

$$p_{prior} = p(w \in p_i) = \frac{PagesContainingWord}{TotalPages},$$

where Total Pages = 247,491. If p_i contains the word w and points to a new page p_2, then the probability of the word w appearing in p_2 is computed as

$$p_{posterior} = p(w \in p_2 | p_i \rightarrow p_2 \wedge w \in p_i)$$

$$= \frac{LinkedPagesContainingWord}{OutgoingLinks}$$

For instance, in the wt2g corpus there are 55,654 pages that contain the word *each*, and these pages have a total of 46,163 links pointing to the pages in the collection, 15,815 of which have the word *each*. Therefore, its prior probability is $\frac{55654}{247491} = .225$, and its posterior probability is $\frac{15815}{46163} = .343$.

We are interested in the ratio of posterior over prior probability for each stemmed word and would like to see if there is any interesting relationship between this ratio and other linguistic features.

We will look at the ratio $R = p_{posterior}/p_{prior}$ (the *link effect*) which describes how much more likely a linked page is to contain a given word than a random page.

IDF (Inverse Document Frequency) values based on the wt2g corpus are also computed. We compute IDF using the formula $idf(w) = -\log_2(1 - e^{df(w)/nd})$, where $df(w)$ is the document frequency (fraction of all documents containing w) and nd is the number of documents in the collection.

2.1 Results and Discussion

Table 1 shows the different measures for the 2000 words with lowest IDF. Each line shows the average values on a chunk of 100 words.

As one can see in the table, the posterior probabilities are always higher than the prior probabilities. Hypothesis testing shows that the difference between prior and posterior is statistically significant, which verifies our assumption. It is noticeable that the link effect has the same trend as the IDF values. The correlation coefficient of these two columns is 0.7112. It is customary to use IDF as an indicator of words' content. Low IDF usually implies a low content value. We would like to investigate whether link effect can be used instead of IDF for certain IR tasks. Let's consider the sample words *between* and *american* on table 1. Intuitively, *american* has more content than *between*, but the later has an IDF of 2.37, higher than that of the former (2.36). However, their link effects agree with intuition: *american*: 1.97, one standard deviation higher than *between*: 1.40.

Table 2 compares the link effects for two ranges of sample words with roughly the same IDF values within each range. It shows the words in the order of IDF and of Link Effect (R). As one can see, the link effect tends to be high for content words when IDF value alone cannot discriminate the words.

$IDF \approx 3.0$				$IDF \approx 4.0$			
sorted by IDF		sorted by link effect R		sorted by IDF		sorted by link effect R	
word	IDF	word	R	word	IDF	word	R
human	2.981	close	1.675	centuri	3.988	extend	2.085
accord	2.983	among	1.770	interact	3.990	beyond	2.477
perform	2.984	further	1.796	introduct	3.993	front	2.606
close	2.985	expect	1.864	front	3.994	centuri	2.713
press	2.992	accord	1.922	travel	3.997	elimin	2.753
applic	2.992	assist	1.962	elimin	4.009	damag	2.757
expect	2.997	human	2.093	opinion	4.013	introduct	2.843
among	2.998	perform	2.095	damag	4.017	opinion	2.984
assist	3.004	applic	2.203	beyond	4.019	travel	3.491
further	3.011	press	2.388	extend	4.021	interact	3.527

Table 2: IDF vs. link effect R

Figure 1 describes a linear fit of R over the 2000 words with the lowest IDF in our corpus. A very clear trend can be observed, whereby over most words, the value of R is almost a constant. When we looked only at the top 100 or 200 words, the trend was even cleaner. However, with 2000 words one cannot help but notice that a number of outliers appear in the left hand part of the figure. We ran a K-Means c (with K=2) to identify two clusters of words. The clusterer stopped after 32 iterations after identifying the two clusters (Figures 2 and 3), each with a very clear trend. Their means are 1.86 and 3.57, respectively.

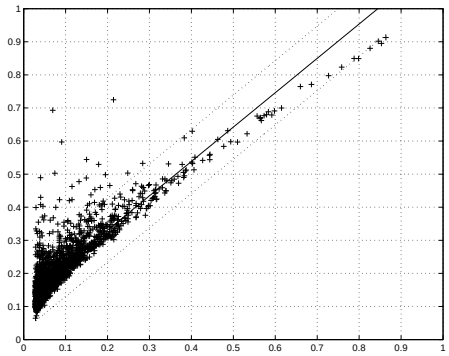

Figure 1: Linear fit for 2000 lowest-IDF words. The X axis represents the prior probability p while the Y axis corresponds to the posterior probability p'.

3 Conclusion

In this paper we discussed some properties of hyperlinked Web documents. We showed that the probability of a word w_j to appear in a Web document d_i depends on the presence of w_j in documents pointing to d_i.

Words	Prior	Posterior	IDF	link effect R	Sample words
1-100	0.4047	0.5293	1.6761	1.3639	the, of, make, and
101-200	0.2141	0.3574	2.3803	1.6745	under, go, between, amlusterererican
201-300	0.1688	0.3209	2.6896	1.9047	market, subject, special, mean
301-400	0.1386	0.2876	2.9513	2.0750	administr, put, establish, ask
401-500	0.1192	0.2588	3.1548	2.1750	understand, social, hand, share
501-600	0.1046	0.2426	3.3326	2.3179	prevent, staff, risk, north
601-700	0.0934	0.2246	3.4879	2.4085	trade, class, size, california
701-800	0.0839	0.2201	3.6354	2.6233	global, drug, letter, softwar
801-900	0.0752	0.2004	3.7884	2.6668	sound, tool, monitor, transport
901-1000	0.0669	0.2024	3.9499	3.0200	permit, target, east, normal
1001-1100	0.0605	0.1823	4.0909	3.0149	approxim, telephon, danger, europ
1101-1200	0.0548	0.1710	4.2292	3.1213	favor, richard, map, pictur
1201-1300	0.0498	0.1752	4.3635	3.5210	professor, earth, english, republican
1301-1400	0.0454	0.1652	4.4934	3.6366	medicin, doctor, church, color
1401-1500	0.0416	0.1630	4.6166	3.9224	permiss, agenda, programm, prioriti
...	...		...		...
100001-100100	0.0000	0.0642	12.4774	363.7331	sinker, surmont, thong, undergrowth
500001-500100	0.0000	0.0215	16.9270	2658.9231	scheflin, schena, schendel, scheriff

Table 1: Some measurements over 20 sets of 100 words among the 2000 lowest-IDF words plus 2 sets of 100 words among the words of higher IDF

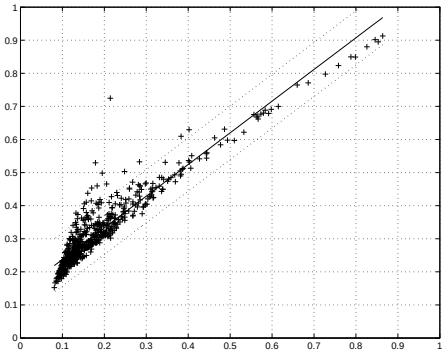

Figure 2: Linear fit for Cluster 1, which contains many low-IDF words such as *by*, *with*, *from*, etc.

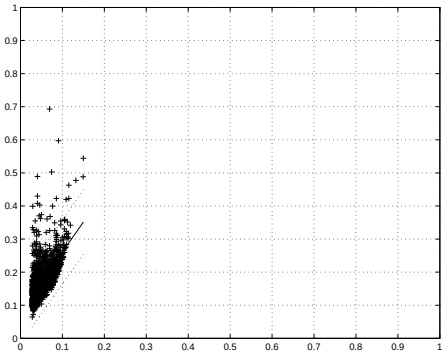

Figure 3: Linear Fit for Cluster 2. Sample words from this cluster are *photo*, *dream*, *path*, etc.

References

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286:509–512.

Abraham Bookstein and Don Swanson. 1974. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):118–132.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *ACL-96*, pages 310–318, Santa Cruz, CA. ACL.

Kenneth Church and William Gale. 1995. Posson mixtures. *Natural Language Engineering*.

Kenneth Church. 2000. Empirical estimates of adaptation: the chance of two noriegas is closer to $p/2$ than p^2. In *COLING*, Saarbruecken, Germany, August.

W. Bruce Croft and David J. Harper. 1979. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295.

Brian Davison. 2000. Topical locality in the web. In *SIGIR 2000)*, Athens, Greece, July.

P. Erdős and A. Rényi. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61.

David Hawking. 2002. Web research collections - trec web track. http://www.ted.cmis.csiro.au/TRECWeb/.

Filippo Menczer. 2001. Links tell us about lexical and semantic web content. http://arxiv.org/cs.IR/0108004.

Frederick Mosteller and David L. Wallace. 1984. *Applied Bayesian and Classical Inference - The Case of The Federalist Papers*. Springer Series in Satistics, Springer-Verlag.

Jay Ponte and Bruce Croft. 1998. A language modeling approach to information retrieval. In *SIGIR 1998*, pages 275–281, Melbourne, Australia, August.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY.