Preparation of Topical Reading Lists from the Link Structure of Wikipedia

Alexander D. Wissner-Gross
Department of Physics, Harvard University
17 Oxford Street, Cambridge, MA 02138 USA
alexwg@physics.harvard.edu

Abstract

Personalized reading preparation poses an important challenge for education and continuing education. Using a PageRank derivative and graph distance ordering, we show that personalized background reading lists can be generated automatically from the link structure of Wikipedia. We examine the operation of our new tool in professional, student, and interdisciplinary researcher learning models. Additionally, we present desktop and mobile interfaces for the generated reading lists.

1. Introduction

Often we do not know what we do not know. This situation is traditionally resolved by having information pushed to us, for example, through textbooks. There has been much recent interest in personalized information delivery through pull-based methods, such as search engine queries [1]. Personalization of push technology [2] also has received attention for, among other applications, news delivery [3], rule-based curriculum preparation using hand coded metadata [4], and pre-trained connectionist-based curriculum preparation [5].

Consider three archetypal future scenarios:

(1) A professional with a broad educational foundation needs to be brought up to speed quickly on a new client's industry. Rather than laboriously reading books to comprehensively patch the knowledge gap, the professional's computer prepares personalized background reading on the industry.

(2) A high school student is captivated by a new model of helicopter hovering in the sky. The student is scheduled to take a physics course soon and the student's camera-equipped phone knows this. Over the next hour, it provides short tutorials on classical mechanics, the history of aerodynamics, and the technical developments leading up to the modern helicopter. The student acquires undergraduate-level knowledge at a precocious age because the education was motivated by and tailored to the environment (after [6]).

(3) An interdisciplinary researcher uses software to prepare a course of study at a rapidly growing interface between fields, such as chemical engineering and biotechnology, for which few authoritative, yet up-to-date, texts exist.

These scenarios are illustrative, respectively, of 'top-down' (from seed topics outwards; see Figure 1A), 'bottom-up' (from broad source topics inward to specialized sink topics; see Figure 1B), and 'horizontal' (from topics related to all seed topics outwards to the seeds themselves; see Figure 1C) topic orderings.

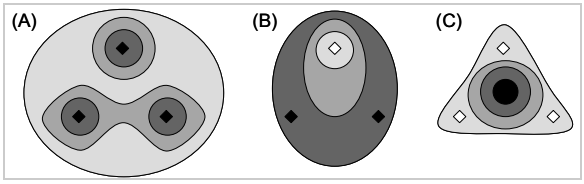

Figure 1. Different topic ordering schemes, with seed topics represented by diamonds and ordering from dark to light.

To enable these distinct scenarios with a single system, a frequently-updated corpus is needed that encompasses most general-use topics. It must also be possible to deduce relationships and ordering of different topics from the corpus, without added human

Proceedings of the Sixth International Conference on Advanced Learning Technologies (ICALT'06)
0-7695-2632-2/06 $20.00 © 2006 IEEE

intervention. Finally, the corpus should be inexpensive, so that it may receive widest possible distribution. The apparent best solution under these constraints is the Wikipedia hypertext [7], whose link structure has already been exploited for semantic automation [8].

It should be noted that Wikipedia already supplies a wide variety of topic lists, called "categories" [9], that are superficially similar to those we describe. However, the categories are not ordered and are not dynamically generated, ruling out direct application to the general case of the above scenarios. Clearly, there is a great deal more semantic information in the natural language text itself of Wikipedia than there is encoded in the link structure. While there have been significant recent advances in the recovery of ontologies from natural language [10], this work will focus on the simplest and most obvious source of topic connectivity information—the link structure—with the expectation that natural language processing capability can be added in future iterations.

2. Top-down ordering

For the top-down scenario, which most resembles traditional search, we implemented the PageRank algorithm [1] with tunable biases for article ordering. A general calculation of PageRank by the power method [11] is performed by iterating until convergence

$$\vec{v} \leftarrow \alpha\hat{A}\vec{v} + (1-\alpha)\left[\beta\vec{e}_{pers} + (1-\beta)\sum_i \frac{\vec{e}_i}{N}\right],$$

where $\vec{v}$ is the vector whose coordinates are the article rank values, $\hat{A}$ is the Markov matrix of the link graph of Wikipedia, $\vec{e}_{pers}$ is the personalization unit vector whose coordinates have a uniform nonzero value if they correspond to seed topics and zero otherwise, $\vec{e}_i$ is the unit vector along article i, $\alpha \equiv d/(1+d)$ parameterizes mean reachable article distance d from seed articles (assuming no back-links), and β parameterizes the degree of personalization.

In the original PageRank, $\beta = 0$. Although $\beta < 1$ could conceivably allow useful corpus-wide influence on the topic listing, highly linked and often irrelevant "attractor" articles tended to be promoted in practice (see Table 1) and so $\beta = 1$ was ultimately used.

The other parameter, α, may be varied according to desired list size and breadth. For $\alpha << 1$, the ordering is determined primarily by the shortest path lengths from the seed articles. Increasing α increases the symmetry-breaking influence of inbound links. The mean log outdegree of Wikipedia was found to be $\langle \ln b \rangle = 1.17 \pm 1.32$. Therefore, for preparation of a list of $N = 10$ articles, a mean depth of

$$d = \log_b\left[N(b-1)+1\right] - 1 \approx 1.69,$$

corresponding to $\alpha \approx 0.63$, was used.

3. Implementation

The June 23, 2005, snapshot of the English Wikipedia articles without images (2.8 GB in size) was used as a link corpus. Link structure was extracted by regular expression matching from the raw SQL file. Iteration was halted after at most $15N$ steps (where $N \sim 1.4\times10^6$ is the number of articles), after the observation that 15 iterations are generally needed for PageRank convergence [11]. The process completed in several minutes. Meta-articles, such as author pages, were discarded prior to ranking.

Table 1. Influence of personalization tuning parameter, , on top-down ordering for "Helicopter" example, with *d*=0.1. Topics covered by maximum personalization are underlined.

Order	β = 0.01	β = 0.1	β = 1.0
1	Helicopter	Helicopter	Helicopter
2	2000	United States	20th century
3	United States	2000	U.S. Navy
4	Race (U.S. Census)	U.S. Navy	Igor Sikorsky
5	Wikipedia	20th century	Jan Bahyl
6	United Kingdom	Jan Bahyl	Kamov Ka-50
7	Marriage	Igor Sikorsky	Robinson Helicopter
8	U.S. Census bureau	United Kingdom	United Kingdom
9	England	Kamov Ka-50	United States
10	Asian (U.S. Census)	Wikipedia	Westland Aircraft

4. Bottom-up and horizontal orderings

Thus far, we have discussed the top-down scenario, in which a quasi-breadth first ordering out from selected seed topics is desired. Although seed topics of differing weights could be accommodated, the remaining two scenarios require further processing of the ranks described in Section 2 to yield useful orderings.

As shown in Figure 2, for distances below 6 links, the calculated rank of articles will decrease approximately exponentially with distance from a seed article due to a relatively constant branching factor. Therefore, in the bottom-up scenario, to linearly parameterize the ordering from source topics (e.g., "physics") to sink topics ("helicopter") of an article *k*, the ratios,

$$\prod_i v_{\text{source}(i)\to k} \Big/ \prod_j v_{\text{sink}(j)\to k} ,$$

of source rank to sink rank are calculated. As a relevance cutoff, the product,

$$\prod_i v_{\text{source}(i)\to k} \cdot \prod_j v_{\text{sink}(j)\to k} ,$$

of the source and sink ranks is applied to achieve the desired list size. Note that this approach generalizes trivially to multiple sources and sinks. For comparison, graph distances from seed topics are calculated, ordered by the difference of distances,

$$\sum_{i,j} \left| d_{\text{source}(i)\leftrightarrow k} - d_{\text{sink}(j)\leftrightarrow k} \right| ,$$

and cut off by the sum of distances,

$$\sum_i d_{\text{source}(i)\leftrightarrow k} + \sum_j d_{\text{sink}(j)\leftrightarrow k} ,$$

as demonstrated in Table 2. For short lists, distance ordering is found to be more robust against the inclusion of less relevant, highly linked topics. For both merged PageRank and distance ordering, the lists are unbalanced between the seed topics due to different numbers of closed loop self-references and different seed topic out-degrees, respectively.

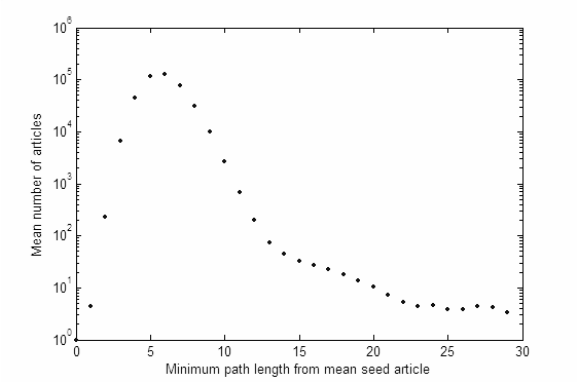

Figure 2. Average number of articles at different minimum path lengths from a seed article (sample size of 100).

Table 2. Comparison of methods for bottom-up ordering example (from "Physics" to "Helicopter"). Minimum path lengths are indicated for the distance ordering.

Order	Merged PageRank	Distance ordering (Source:Drain)
1	Physics	Physics (0:2)
2	Greek language	Astronomy (1:2)
3	20th century	20th century (1:1)
4	United States	Greek language (1:1)
5	Italy	1950s (2:1)
6	United Kingdom	United States (2:1)
7	1950s	15th century (2:1)
8	1916	1961 (2:1)
9	China	1946 (2:1)
10	Helicopter	Helicopter (3:0)

For the horizontal scenario, orderings are

Table 3. Comparison of methods for horizontal ordering example (at the interface of "Chemical Engineering" and "Biotechnology").

Order	Merged PageRank	Distance ordering (Drain 1:Drain 2)
1	19th century	Ecotourism (2:2)
2	X-ray	United States (2:2)
3	United Kingdom	New Zealand (2:2)
4	Latin	Australia (2:2)
5	2003	Biotechnology (1:0)
6	2000	Convention on Biological Diversity (2:1)
7	World War II	E. coli (2:1)
8	Germany	United Nations (2:1)
9	2002	DNA microarray (2:1)
10	Soviet Union	Genetic engineering (2:1)
80	Convention on Biological Diversity	Republic of Congo (3:2)
81	Radiation	Kenya (3:2)
82	Enterobacteriaceae	Seychelles (3:2)
83	Genetic engineering	Swaziland (3:2)
84	Salmonella typhi	Guinea-Bissau (3:2)

parameterized by absolute differences of logs of ranks,

$$\sum_{i,j}\left|\log v_{\text{seed}(i)\to k} - \log v_{\text{seed}(j)\to k}\right|,$$

and absolute differences of distances,

$$\sum_{i,j}\left|d_{\text{seed}(i)\leftrightarrow k} - d_{\text{seed}(j)\leftrightarrow k}\right|,$$

with analogous cutoff methods. In this scenario, where long curricula may be desirable, distance-based ordering is at a disadvantage because large numbers of similar and unrelated articles can lie at identical distances from the seed topics (e.g., a set of countries equidistant from the topics in Table 3). While PageRank-based orderings initially promote unrelated attractor articles, they eventually resolve to related articles (see Table 3).

5. Interfaces

The three scenarios described necessitate two different interfaces. As a desktop interface for the top-down and horizontal scenarios, custom RSS feeds can be generated (see Figure 3). For the bottom-up scenario, a mobile interface is desirable. Notably, mass storage capacities for many wearable devices, such as the Apple iPod and hard drive-based cell phones, have exceeded the Wikipedia text corpus size. Therefore, mobile generation and presentation (see Figure 4) of algorithmically generated curricula is now possible. Better-readable mobile displays and advances in mobile visual search will still be needed to complete the bottom-up learning scenario.

Figure 3. RSS feed for "Condensed matter physics" top-down ordering in Mozilla Thunderbird.

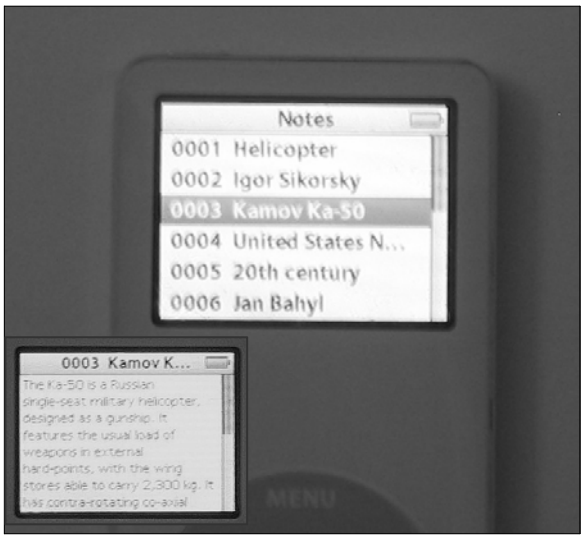

Figure 4. Automatically generated top-down curriculum on the "Helicopter" for the Apple iPod nano, presented with the built-in Notes applet. (Inset) An article from the reading list.

6. Conclusions and future work

We have reported a novel corpus-algorithm combination for dynamically generating personalized background reading lists under various usage scenarios, using the link structure of Wikipedia and PageRank and graph distance-based ordering. We also presented desktop and mobile interfaces for the generated reading lists.

Much optimization work remains for topic ordering. In particular, it might be constructive to hybridize the robustness of distance-based ordering with the graph structural sensitivity of PageRank-derived techniques. Our technique discards a great deal of semantically valuable information, including category membership and the natural language text itself, which if exploited properly might significant improve list quality. Finally, the interfaces require full usability testing and preliminary tests are underway. Nonetheless, dynamic preparation of reading lists, for mobile consumption in particular, is a tantalizing prospect.

Acknowledgements

The author gratefully acknowledges financial support by the Fannie and John Hertz Foundation.

References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical report, Stanford University, 1998.

[2] M. Franklin, S. Zdonik, "'Data in Your Face': Push Technology in Perspective", *Proc. SIGMOD 1998*, ACM Press, 1998, pp. 516-519.

[3] P.R. Chesnais, M.J. Mucklo, and J.A. Sheena, "The Fishwrap personalized news system", *Proc. 2nd International Workshop on Community Networking*, IEEE Press, 1995, pp. 275-282.

[4] O. Conlan, V. Wade, C. Bruen, and M. Gargan, "Multi-model, Metadata Driven Approach to Adaptive Hypermedia Services for Personalized eLearning", *Proc. AH 2002*, Springer-Verlag, 2002, pp. 100-111.

[5] K.A. Papanikolaou, G.D. Magoulas, and M. Grigoriadou, "A Connectionist Approach for Supporting Personalized Learning in a Web-Based Learning Environment", *Proc. AH 2000*, Springer-Verlag, 2000, pp. 189-201.

[6] J. Smart, "Simulation, Agents and Accelerating Change", *Accelerating Change 2004*, Acceleration Studies Foundation, 2004.

[7] J. Voss, "Measuring Wikipedia", *Proc. ISSI 2005*, in press.

[8] D. Grangier and S. Bengio, "Inferring document similarity from hyperlinks", *Proc. CIKM 2005*, ACM Press, 2005, pp. 359-360.

[9] T. Holloway, M. Bozicevic, and K. Börner, "Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors", *arXiv:cs.IR*/0512085.

[10] P. Cimiano, A. Hotho, and S. Staab, "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis", *J. Artif. Intell. Res. 24*, 2005, pp. 305-339.

[11] S. Kamvar, T. Haveliwala, and G. Golub, "Adaptive methods for the computation of PageRank", *Lin. Alg. Appl. 386*, 2004, pp. 51-65.