An Unsupervised Ranking Method Based on a Technical Difficulty Terrain[*]

Shoaib Jameel, Wai Lam
Systems Engineering and
Engineering Management,
The Chinese University of
Hong Kong
{msjameel,wlam}@se
.cuhk.edu.hk

Ching-man Au Yeung[†]
ASTRI,
3/F Bio-informatics Centre, 2
Science Park West Avenue,
Hong Kong
albertauyeung@astri.org

Sheaujiun Chyan
Systems Engineering and
Engineering Management,
The Chinese University of
Hong Kong
chyan.sheaujiun@gmail.com

ABSTRACT

Users look for information that can suit their level of expertise, but it often takes a mammoth effort to trace such information. One has to sift through multiple pages to look for one that fits the appropriate technical background. In this paper, a query-independent ranking system is proposed for technical web pages. The pages returned by the system are sorted by their relative technical difficulty in either ascending or descending order specified by the user. The technical difficulty of a document i.e. terms in sequence, is first computed by the combination of each individual term's geometry in the low-dimensional latent semantic indexing (LSI) space, which can be visualized as a conceptual terrain. Then the pages are ranked based on the expected cost to get over the terrain. Results indicate that our terrain based method outperforms traditional readability measures.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Retrieval models, Search process, Selection process; H.3.7 [**Digital libraries**]

General Terms

Algorithms, Experimentation

Keywords

Conceptual Hop Model, LSI, Technical Expertise

1. INTRODUCTION

Readers can find it difficult to comprehend technical articles outside their domain of specialization because they are not acquainted with the jargon or technical concepts of that domain. A computer scientist looking for introductory information about "fever" will probably want to visit a web page meant for general public. A medical research student would probably want to go to websites meant for medical researchers in order to know the current trends in pharmaceutical research. Search engines currently return a mix of introductory and advanced documents [8], which means that a person has to carefully read the web pages one by one in order to stumble upon the one that can fit her background knowledge and technical understanding. We propose a ranking system that can sort web pages based on technical difficulty for any domain. We evaluate the model on the medical domain.

Domain expertise relates to the knowledge of the subject of a particular domain [15]. In our work, technicality of each term is computed based on the observation in the LSI space. Then, we apply this formulation towards web page ranking where the "Conceptual Hop Model" (CHM) follows term order in the web page and "hops" from one term to another in succession which incurs certain "cost". In the end, an expected cost is computed. If the expected cost is high, the web page is more technical. To make text comprehensibility easy for the learner, the path that the learner has to face during reading process should be as easy as possible [12].

The main contributions of this work are: a new method to find the technicality of every term in the web page in the low dimensional latent semantic space and a new method which computes the "cost" involved when hopping from one term to another in the terrain.

2. RELATED WORKS

Traditional readability formulae have been used extensively in determining the grade level of articles. Flesch-Kincaid [7], Dale-Chall[1] and Lexile$^{\mathrm{TM}}$[14] etc. are some of the commonly used readability measurement techniques. Wolfe et al. [16] describe the use of latent semantic analysis (LSA) in matching readers and texts. Textual coherence

[*]The work is supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Codes: CUHK4128/07 and CUHK413510). This work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

[†]Au Yeung contributed to this paper while he was at the NTT Communication Science Laboratories, Kyoto, Japan.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

Table 1: A TDM built by selecting terms from PubMed and Yahoo! Health documents.

	P1	P2	P3	P4	Y1	Y2	Y3	Y4	Y5
acne	1	0	0	0	1	0	0	0	0
insulinotropic	1	0	0	0	0	0	0	0	0
postrandial	1	0	0	0	0	0	0	0	0
abnormal	0	1	0	0	0	0	1	0	0
ichthyosis	0	1	1	1	0	0	0	0	0
mitotic	0	1	1	0	0	0	0	0	0
mutation	0	0	1	1	0	0	0	0	0
found	0	0	0	1	0	1	0	0	0
human	0	0	0	1	0	1	0	0	0
follicle	0	0	0	0	1	0	0	0	0
later	0	0	0	0	1	1	0	0	1
body	0	0	0	0	0	1	0	0	0
death	0	0	0	0	0	1	0	0	1
oncogenes	0	0	0	0	0	1	1	2	0
result	0	0	0	0	0	0	1	1	0
transformed	0	0	0	0	0	0	1	0	1
cancer	0	0	0	0	0	0	0	1	1

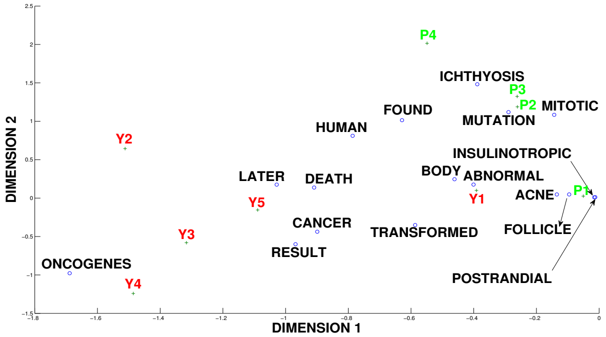

Figure 1: A two-dimensional plot showing the position of terms (circular markers) and documents (plus shaped markers) in the latent semantic space.

is another related concept that uses statistical or natural language processing techniques to determine the text comprehensibility [5, 4].

In [8], the authors begin with the definition of the *introductory* and *advanced* documents. They have used a supervised learning method to build a classifier (FAMCLASS) to predict web pages as introductory or advanced. In [2], the authors build a smoothed unigram model to predict the grade level of the text. The authors have stated that traditional readability measures perform poorly on web pages because they are short, contain lots of noise (like irrelevant strings due to pre-processing), presence of tables, images etc.

Nakatani et al. [13], describe a method for re-ranking search results of a web search engine (Easiest-First Search) in descending order of their comprehensibility using the Japanese Wikipedia. The work of Yan et al. [17] has similar objective as in this paper. The authors state that document scope and document cohesion are important parameters in finding simple texts. The authors have used a controlled vocabulary thesaurus termed as Medical Subject Headings (MeSH). However, the applicability of such methods in the other domains will always be a concern as one requires similar kind of controlled vocabulary. In our model, CHM, distance between consecutive terms is a measure of the "conceptual cohesion" between terms and "hops" between terms gauge the difficulty in conquering that bigram.

3. OBSERVATIONS IN THE LATENT SEMANTIC SPACE

Latent Semantic Indexing [3] makes use of the Term Document Matrix (TDM) on which a matrix decomposition method "Singular Value Decomposition" (SVD) is applied. The most noticeable observation is that terms which are unique or central to a document remain close to the document vectors. Terms which are used frequently, e.g. the general English language words like *"result"*, *"human"*, *"body"* etc., are usually far away from the highly technical document vectors. Only few technical terms move far away from their document vectors.

We construct a small document collection by selecting sentences from Yahoo! Health (document IDs beginning with alphabet Y) and PubMed (document IDs beginning with alphabet P) for illustrating our idea. These terms have been taken from sentences present in different documents from our real test collection, for example one such sentence from PubMed corpus is - *"**Mitotic** recombination in patients with **ichthyosis** causes reversion of dominant **mutations** in KRT10."*, where the terms in bold have been considered in the TDM. The choice of terms in the TDM is based on the property of the corpus itself. For instance, PubMed articles are more technical than Yahoo! Health, therefore more technical terms have been chosen from PubMed than Yahoo! Health. SVD is then applied on the TDM to get the term and the document co-ordinates in two dimensions. A similar small example can be found in [10]. Figure 1 is a reconstruction of the original vector space after reducing the dimensions to two by applying SVD to the TDM in Table 1.

From Figure 1 it is evident that PubMed and Yahoo! Health articles form two separate clusters in the latent space. Technical terms like *"follicle"*, *"postrandial"* etc., are very close to their own document vectors. General terms like *"body"*, *"found"*, *"transformed"*, *"human"*, *"result"* etc., can be seen far from their document vectors. $Y1$ and $P1$ describe about the same theme *"acne"*, hence are close to each other. But $P1$ being more technical, has many technical terms in close proximity. Technical terms are closer to $P1$ than $Y1$, suggesting $P1$ is more technical than $Y1$.

3.1 The Notion of Term Technicality

Similarity between a document and a term in the latent semantic space is inversely proportional to the distance between them [9]. Based on the concept of similarity, "technicality" is the amount of technical tinge a term gives to the document and is based on the closeness to the document vector in the low dimensional latent semantic space.

Mathematically, the technicality F_t^d of a term t in document d can be written as:

$$F_t^d = \frac{idf_t}{(r_t^d + \epsilon)} \tag{1}$$

where, r_t^d is the semantic distance (Euclidean) between the term and document vectors in the latent semantic space. ϵ is a small constant added to avoid the case when $r_t^d = 0$. idf_t is the Inverse Document Frequency of term t.

Consider the term *"later"* in Figure 1. This is a general term but is located close to the document vector $Y5$ in the latent space. In a large document collection, the global importance of this term will be less. Hence, idf_t ensures that technicality of a general term remains low. *"oncogenes"* is

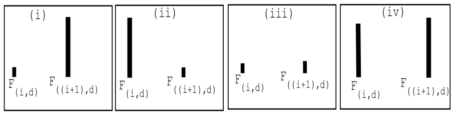

Figure 2: Four important scenarios when transiting from one term to another in the terrain.

shared among three documents. It is a technical term and hence the global importance will be high. It can be clearly seen from the TDM (Table 1) that $Y2$ has more general terms than $Y4$. Hence, $Y4$ is more technical than $Y2$ (and also $Y3$). Therefore, *"oncogenes"* is more central to $Y4$. Technical term like *"follicle"* is closer to the document vector and its $idf_{(follicle)}$ will be high, hence technicality will be high. General terms like *"found"*, *"human"* etc. will have low technicality as their idf_t will be low and their distance from the document vectors will be large in which they occur.

4. THE CONCEPTUAL HOP MODEL (CHM)

4.1 Formulation

Hops in CHM occur by taking two consecutive terms at positions i and $(i+1)$ i.e. a bi-gram in the document d. In doing so, some cost $C^d_{i,i+1}$ is incurred. Let $s^d_{i,i+1}$ denote the "Euclidean distance" between terms in the bi-gram in the latent semantic space. Let $n^d_{i,i+1}$ be the number of "same bigram" that have been encountered previously in the web page d until the current position $(i+1)$.

The cost $C^d_{i,i+1}$ incurred when hopping from term i to term $(i+1)$, which are characterized by technicalities F^d_i and F^d_{i+1} respectively in document d, separated by a distance $s^d_{i,i+1}$ is given as:

$$C^d_{i,i+1} = \frac{(F^d_i + F^d_{i+1}) \times [s^d_{i,i+1}]^{sgn(s^d_{i,i+1}-1)\times F^d_{i+1}}}{n^d_{i,i+1}+1} \qquad (2)$$

where $sgn(...)$ is defined as the signum function that extracts the sign of a real number.

4.2 Analysis

Figure 2 shows four important cases in hopping, and will be used as a graphical depiction in order to justify the expressions in Equation 2. The distance $s^d_{i,i+1}$ has been kept constant in all the cases for better understanding.

$(F^d_i + F^d_{i+1})$ measures the combined technicality of the bigram. The limitation of this expression is that it fails to distinguish cases in Figure 2(i) and 2(ii). Cohesion property in the conceptual terrain is handled by $[s^d_{i,i+1}]^{sgn(s^d_{i,i+1}-1)\times F^d_{i+1}}$. If the terms at positions i and $(i+1)$ come from completely different concept spaces, difficulty in understanding them will be high as the reader encounters a "big conceptual leap" in order to relate two completely different concepts. There might be cases where the Euclidean distance between the term pairs may be $0 < s^d_{i,i+1} < 1$ which will affect the monotonicity of the function. Thus the expression $sgn(s^d_{i,i+1} - 1)$ is introduced. A document contains more number of non-domain concepts than domain specific concepts. Consequently, the cost values for such documents will be much skewed. In order to make the cost computation more sensitive to different technicalities, F^d_{i+1} has been chosen as an exponent to compensate for the skewness. If a reader has already encountered the same bigram before; then the ease with which a reader will cross the same bigram in the next encounter will be relatively more than that in the previous encounters. Division by the number of same bigrams, $n^d_{i,i+1}$ seen so far ensures that there is a constant reduction in cost in every same encounter.

5. THE RANKING FUNCTION

If the terrain is highly rugged or the document is too technical; then the reader will face immense difficulties in comprehending it. The ranking function considers the expected cost in order to complete the terrain. The expected cost E^d for document d can be written as:

$$E^d = \frac{\sum_{i=1}^{W_d - 1} C^d_{i,i+1}}{W^d - 1} \qquad (3)$$

where $C^d_{i,i+1}$ is defined in Equation 2 and W^d is the number of terms in document d.

6. EXPERIMENTS AND RESULTS

6.1 Data Set - Medical Corpus

Medical documents contain a high distribution of technical concepts. We crawled web pages from two different sources, Yahoo! Health (2976 documents with 20109 unique terms) and PubMed (3087 documents with 145016 unique terms). Yahoo! Health articles are written for general public who have little or no understanding of medical concepts. PubMed documents are research papers written by health professionals.

6.2 Comparison with the Baselines

Most of the procedures for experimentation were followed from [2]. Hence, major details about the baselines (the baselines are Models: M3 and M4 in Table 2) described there are being omitted here. One change that is made relates to the number of words used, where in [2] the authors have used 100 token passages but we use full length passages and then normalize by the total number of words in the document. CHM's semantic components are "technicality" and "conceptual cohesion" with no syntactic component.

Flesch-Kincaid readability metric with both semantic and syntactic components is considered in our experiments. SMOG [11] with its semantic and syntactic components is another baseline. To compute these measures "style"[1] package was used. We add all the technicalities for a web page (computed in Section 3.1) and take its mean (Model $M7$). Hence, term order is lost in this baseline. Model $M6$ is another baseline where we remove the technicalities from the CHM and use the $tf \times idf$ values instead.

6.3 Human Ratings for top-k Pool

Normalized Cumulative Discounted Gain (nDCG) [6] was chosen as the performance measure, and is calculated as:

$$W(i) = \frac{1}{Z_n} \sum_{i=1}^{n} \frac{2^{r(i)} - 1}{\log(1+i)} \qquad (4)$$

where Z_n is the normalization constant such that a perfect list gets a score of 1, $r(i)$ denotes the rank label of the i^{th} document in the ranked list, n is length of the ranked list.

[1] *http : //www.gnu.org/software/diction/diction.html*

Table 2: nDCG@k for Medical corpus.

Reference Numbers	Models	nDCG@10	nDCG@50	nDCG@100	nDCG@150	nDCG@200
M1	CHM without stop words	**1.0**	**1.0**	**1.0**	0.997	0.997
M2	CHM with stop words	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
M3	MLF	0.066	0.066	0.075	0.077	0.075
M4	Flesch-Kincaid	0.292	0.376	0.410	0.424	0.430
M5	SMOG	0.215	0.299	0.348	0.371	0.388
M6	CHM-$tf \times idf$ with stopwords	**1.0**	**1.0**	**1.0**	0.988	0.965
M7	Using the technicality mean	**1.0**	0.984	0.963	0.958	0.962

"depth-k" pooling was adopted where "top-k" documents from all models were accumulated and the rest were assumed irrelevant. In the "top-k" pool where $k = 200$, the number of unique documents were 1249. Annotations from three people were used (mode of the scores was considered) who gave the relevance rating values. Two of them had basic knowledge of medical science while one was a medical student.

Relevance rating is in the range of "0 and 4". "0" means that the document is completely advanced at that position. "1" indicates that the document contains rich use of jargon or technical terms without explaining them. Relevance rating of "2" means that the text requires some background knowledge of the topic because some technical terms are not explained. "3" means that the author has used technical terms and has defined them. Relevance rating of "4" means that the document is technically simple without the use of any jargon and the text can be easily comprehended by a beginner looking for technically introductory articles.

Table 2 shows the results when the documents are ordered from technically introductory to advanced. CHM with stop words (Model $M2$) has fared well as compared to the other models as it ranked technically introductory web pages at the top of the results containing no jargon. Model $M1$ ranked some of PubMed's case study articles which contained jargon. Other models ($M3$, $M4$ and $M5$) mainly tap readability features but CHM is able to capture technical difficulty i.e. technical concepts and the cohesion between concepts. These models ($M3$, $M4$, $M5$ and also $M6$, $M7$) ranked few medical news and quiz web pages at the top of the results, which indeed were not meant for a beginner who wanted to read technically simpler articles.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the problem of determining the technical difficulty of a web page in domain specific information retrieval. Ranking documents based on technical expertise can help in learning and also development of reader's expertise [15]. Different readability formulae are not directly applicable even though their syntactic and semantic components are used. CHM is able to capture word level technical difficulty, concept level cohesion, and the intricacy experienced when hopping from one term to another in the concept terrain. In future, importance of hyperlinks would be studied, which act as an "added help" given by the author to the reader. If a technical term has been hyperlinked to another web page, the author expects the reader to go to that web page, get acquainted with the definition of the term and come back to the original web page.

8. ACKNOWLEDGMENTS

The authors would like to thank Wei Gao, Bo Chen, Binyang Li and Haiqin Yang for their comments in the paper.

9. REFERENCES

[1] J. S. Chall and E. Dale. Readability revisited: the new dale-chall readability formula. 1995.

[2] K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. *J. Am. Soc. Inf. Sci. Technol.*, 56:1448–1462, November 2005.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[4] C. R. Fletcher, S. T. Chrysler, P. van den Broek, J. A. Deaton, and C. P. Bloom. The role of co-occurrence, co-reference, and causality in the coherence of conjoined sentences. In *R. F. Lorch, and E. J. O'Brien (Eds.), Sources of coherence in reading*, pages 203–218, 1995.

[5] P. W. Foltz, W. Kintsch, and T. K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Process*, 15:285–307, 1998.

[6] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR*, pages 41–48, 2000.

[7] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Feb. 1975.

[8] G. Kumaran, R. Jones, and O. Madani. Biasing web search results for topic familiarity. In *Proceedings of CIKM*, pages 271–272, 2005.

[9] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[10] T. K. Landauer, P. W. Foltz, and D. Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284, 1998.

[11] G. H. M. Laughlin. Smog grading-a new readability formula. *Journal of Reading*, 12(8):pp. 639–646, 1969.

[12] D. S. McNamara, E. Kintsch, N. B. Songer, and W. Kintsch. Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14(1):pp. 1–43, 1996.

[13] M. Nakatani, A. Jatowt, and K. Tanaka. Easiest-first search: towards comprehension-based web search. In *Proceeding of CIKM*, pages 2057–2060, 2009.

[14] A. Stenner, I. Horabin, D. Smith, and M. Smith. The lexile framework. 1988.

[15] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the WSDM*, pages 132–141, 2009.

[16] M. B. W. Wolfe, M. E. Schreiner, B. Rehder, D. Laham, P. W. Foltz, W. Kintsch, and T. K. Landauer. Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, 25(2/3):309–336, 1998.

[17] X. Yan, D. Song, and X. Li. Concept-based document readability in domain specific information retrieval. In *Proceedings of CIKM*, pages 540–549, 2006.