University of Padova at TREC-10

Franco Crivellari Massimo Melucci

Università di Padova
Dipartimento di Elettronica e Informatica

October 28, 2001

1 Introduction

This is the second year that the University of Padova participates to the Text Retrieval Conference (TREC). Last year we participated as well to this program of experimentation in information retrieval with very large document databases. In both years, we participated to the Web track, and specifically to the ad-hoc task, which consists in testing retrieval performance by submitting 50 queries extracted from 50 respective topics. This year we participated to the homepage finding task as well. This year we could devote more time to experiments than last year, yet some problems still arose because we indexed the full-text of documents, while we indexed only a portion of documents only.

2 Approach and Experimental Objectives

This year we participated to the ad-hoc and the homepage finding tasks of the Web track. Our objectives were to evaluate:

1. the effectiveness of passage retrieval in Web page retrieval and homepage finding,
2. the effectiveness of combining classic vector space similarity measure and PageRank measure using all links, and
3. the selection of links of some given types in the previous combination.

The baseline was given by document retrieval based on classic vector space similarity, both for the ad-hoc and the homepage finding tasks. The baseline results served as input to combine themselves with link information. Specifically, the runs being reported in Tables 1 and 2 have been performed. To extract text from Web documents, we employed a software agent that follows the Web links to retrieve the Web pages. This robot has been developed within the National

InterData research project [1]. For the purposes of the TREC experiments, a different version of the robot has been designed and developed because the data to be retrieved were locally stored, and not on the Web. Moreover, the data are encoded in SGML also and then the tool has been modified to deal with this additional format. To only extract the tagged text, our robot employed a tool for HTML syntax analysis, called Tidy, that is reported in [2]. Tidy allows for correcting HTML syntax by adding, for example, missing end tags. Documents have been fully indexed, i.e. all the full-text of each document has been processed to extract keywords and the individual positions at which each keyword occur has been recorded. A stoplist including common Web words, such as web, html, http, com, edu has been used to filter function words out. Words have been stemmed using the Porter's algorithm, yet the original word has been recorded as well. At retrieval time, both individual keywords and keyword pairs have been used. All the performed runs employed a variation of the classic $tf \times idf$ weighting scheme, as expressed below:

$$w_{ij} = (1 + \log tf_{ij})\ (\log \frac{N+1}{n_i})$$

where w_{ij} is the weight of keyword i in document or passage j, t_{ij} is the frequency of keyword i in document or passage j, n_i is the number of documents or passages including i, N is the total number of documents or passages.

A few notes about passage retrieval: In PR runs, a list of 10,000 passages were retrieved in response to the query. The retrieved passage list was then transformed into the corresponding 1,000 documents by summing the respective scores. (The passage list was then transformed into the corresponding 100 documents in case of the homepage finding task.) The more the document includes retrieved passages and the higher these passages are ranked, the higher the document is ranked. Passage size was fixed at 100 words. No formatting or logical structure were used because of the nature of data that made hypotheses on the quality of data a very hard task.

As regards the EP task, note that the same algorithms used for the ad-hoc task were employed. Then, the entry point topics were used as usual queries without any sophisticated processing.

PageRank values were computed for every page by considering all the incoming links up to 10 steps and damping factor at 0.85. The linear function used to combine PageRank values and classic VSM RSVs is $\alpha r + (1-\alpha)v$ where $\alpha = 0.5$, r is a PageRank value and v is a VSM RSV (α was set to 0.5 for the TREC experiments.)

A more detailed illustration is necessary to describe the experiments that tested the effectiveness of selecting links of some given type. Link semantics can enhance link-based retrieval or focused retrieval design. From the one hand, some link-based retrieval algorithms have recently been proposed, e.g. HITS or PageRank to simulate navigation being carried out by end users [3, 4]. Past experiments at TREC have not shown significant effectiveness improvements over the baselines. Probably, there are many noisy links and link filtering algorithms can be enhanced to consider types and filter noisy links out. We use two link types differing on the subgraph topology which they belong to. Such a link points to a given graph topology, independently on the node content. These links are likely to represent organizations of topics accordingly to the given structure, e.g. sequence or tree.

- A sequence link points to a sequence of pages. If the author(s) organize topics in a page sequence then topics are likely to be organized sequentially. We call (x, y) n-sequential link if points to a sequence of n pages. Note that, if (p_0, p_1) is a n-sequential link pointing to $(p_1, ..., p_n)$, then (p_0, p_1) is a 1-sequential link and (p_1, p_2) is a $(n-1)$-sequential link.

- A tree link points to page trees, i.e. page networks without cycles. We call (x, y) n-tree link if it points to a tree being rooted at y with minimum depth n from the root y. Note that, if (p_0, p_1) is a n-tree link, then (p_0, p_1) is a 1-tree link and there exists a $(n-1)$-tree link $(p_1, p_i), i \neq 1$.

Figure 1 depicts an example of 2-tree link (left) and of non-tree link (right), because of a cycle. Figure 2 depicts an example of sequence links. We report some descriptive statistics on structure links. Test data consist of $1,692,096$ full-text Web pages, $1,532,012$ pages with in-going links, $1,295,841$ pages with out-going links, 5.27 in-going links per page, 6.22 out-going links per page. All links are between pages that belong to the test collection; this means that no link points to a page, nor are pages pointed to by links starting outside the collection. Table 2 reports the distribution of sequence and tree links at different values of n, where n is defined above. Note that the percentages of structure, or tree links, out of the total number of in-going or out-going links vary. Most of the out-going links (92.5%) are n-tree links ($n < 10$), and 89.2% of the out-going links are 1-tree or 2-tree links. Only 7.1% out the out-going links are sequence out-going links. A small part of out-going links are not sequence nor tree links; for example, they may point to highly connected graphs. The large majority of structure in-going links are sequence links, yet they are a minority of the

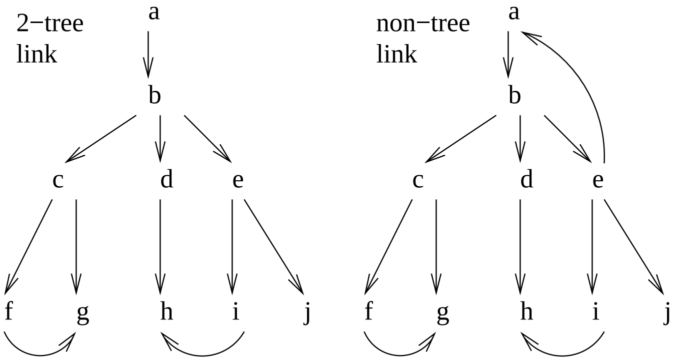

Figure 1: An example of tree and non-tree link.

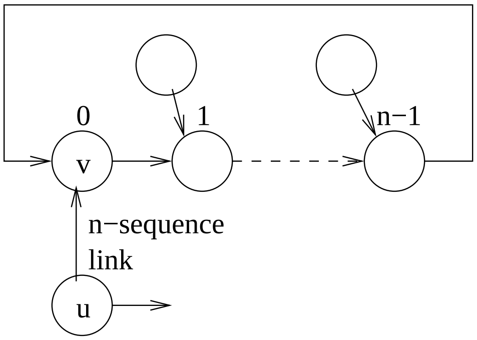

Figure 2: An example of sequences link.

set of in-going links (42.8%). The most apparent result from this preliminary experiment is that a link is likely to point to pages being entry points of small trees of depth 1 or 2. This means that the employed test sample of the Web is a sort of forest of many small trees. It is then likely that page contents are organized accordingly hierarchical structures. Further investigation would be needed to study the topology of these small trees and the relationship between sequence and tree links on the computation of estimates of the popularity and then the relevance of Web pages. The results that might be obtained can be used to enhance link-based retrieval algorithms.

3 Official Results

Table 4 and 5 report the summary of the official results for the ad-hoc task. DR performed better than any other run since 24 out of 50 topic resulted not below the median, while the other runs are below the median for many topics. This means that:

1. passage retrieval performed badly,
2. the combination of PageRank and classic vector space model gave no improvements,
3. selecting tree links gave no improvements in combining PageRank and classic vector space model.

4 Unofficial Results

PDWTAHSL and PDWTEPSL gave no significant variations with respect to other content-link runs.

References

[1] F. Crivellari and M. Melucci. Awir: Prototipo di un motore di ricerca per la raccolta, indicizzazione e recupero di documenti web sulla base dei loro frammenti. Rapporto tecnico T2-S12, Progetto INTERDATA - MURST e Università di Padova: "Metodologie e tecnologie per la gestione di dati e processi su reti Internet e Intranet". Tema 2: "Estrazione di informazioni distribuite sul WWW"., ftp://ftp-db.deis.unibo.it/pub/interdata/tema2/T2-S12.ps, Febbraio 1999. In Italian.

[2] World Wide Web Consortium (W3C) HTML Tidy. `http://www.w3.org/People/Raggett/tidy/`, October 2000. Last visited: October 25th, 2000.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998. Reprinted from [5].

[4] J. Kleinberg. Authorative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the World Wide Web Conference*, 1998. `http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm`.

Run Id.	**Run Type**	**Description**	**(Un)official**
PDWTAHDR	content-only	standard vector space document retrieval using: single words and word pairs, no document normalization; documents retrieved against ad-hoc topics (501-550)	official
PDWTEPDR	content-only	standard vector space document retrieval using: single words and word pairs, no document normalization ; documents retrieved against entry-point topics (EP1-EP145)	official
PDWTAHPR	content-only	standard vector space using: single words and word pairs, no document normalization, prior retrieval 100-words passages, selection of 10000 top passages, retrieval of the corresponding documents; documents retrieved against ad-hoc topics (501-550)	official
PDWTEPPR	content-only	standard vector space using: single words and word pairs, no document normalization, prior retrieval 100-words passages, selection of 10000 top passages, retrieval of the corresponding documents; documents retrieved against entry-point topics (EP1-EP145)	official

Table 1: The summary of the performed runs. Legend: PD = Padova University, WT = Web Track, AH = Ad-Hoc topics, EP=Entry Point topics (homepage finding task), PR = Passage Retrieval, WL = Web In-Links: combination of content and pageranks, TL = Tree In-Links: like WL but only tree in-links are used, SL = Sequence In-Links: like WL but only sequence in-links are used

Run Id.	**Run Type**	**Description**	**(Un)official**
PDWTAHWL	content-link	PDWTAHDR is combined with Google pageranks using a linear function; pageranks are computed using the complete WT link file	official
PDWTAHTL	content-link	PDWTAHDR is combined with Google pageranks using a linear function; pageranks are computed using the tree links only discovered from the WT link file	official
PDWTAHSL	content-link	PDWTAHDR is combined with Google pageranks using a linear function; pageranks are computed using the sequence links only discovered from the WT link file	unofficial
PDWTEPWL	content-link	PDWTEPDR is combined with Google pageranks using a linear function; pageranks are computed using the complete WT link file	official
PDWTEPTL	content-link	PDWTEPDR is combined with Google pageranks using a linear function; pageranks are computed using the tree links only discovered from the WT link file	official
PDWTEPSL	content-link	PDWTEPDR is combined with Google pageranks using a linear function; pageranks are computed using the sequence links only discovered from the WT link file	unofficial

Table 2: The summary of the performed runs. Legend: PD = Padova University, WT = Web Track, AH = Ad-Hoc topics, EP=Entry Point topics (homepage finding task), PR = Passage Retrieval, WL = Web Links: combination of content and pageranks, TL = Tree Links: like WL but only tree links are used, SL = Sequence Links: like WL but only sequence links are used

	sequence		tree	
n	in	out	in	out
1	2,109,946	510,826	424,109	3,944,056
2	401,823	14,958	245,626	3,244,879
3	128,442	15,700	31,487	261,701
4	50,840	22,581	6,551	5,150
5	32,552	6,625	1,473	1,284
6	8,315	146	542	322
7	2,716	48	196	103
8	1,038	11	101	43
9	632	0	79	16
10	471	0	60	5
> 10	1985	0	196	0
total	2,738,760	570,895	710,431	7,457,559

Table 3: The distribution of sequence and tree links at different values of n.

Topic Id.	**N.Rel.**	**Best**	**Median**	**TL**	**WL**	**PR**	**DR**
501	62	18	7	4	13	3	14
502	81	18	6	0	8	5	6
503	33	12	6	2	1	1	1
504	18	13	8	1	9	5	9
505	24	17	11	2	8	0	8
506	2	2	1	0	1	0	1
507	17	11	5	0	1	0	1
508	47	16	7	5	4	2	4
509	140	25	18	3	10	5	10
510	39	25	18	1	9	9	14
511	165	21	16	6	10	7	10
512	14	7	4	0	3	1	3
513	58	13	6	6	3	4	3
514	79	17	8	6	8	5	9
515	41	11	6	5	6	7	7
516	30	9	3	1	3	5	4
517	60	15	3	0	2	3	2
518	84	16	2	1	8	3	7
519	149	20	6	3	4	8	5
520	18	6	3	1	2	0	3
521	57	16	1	0	2	0	2
522	6	5	3	0	3	1	3
523	79	19	3	2	2	11	2
524	35	12	2	2	0	1	0
525	41	11	4	0	3	2	3

Table 4: The summary of the official results (501-525).

Topic Id.	**N.Rel.**	**Best**	**Median**	**TL**	**WL**	**PR**	**DR**
526	49	9	3	5	2	5	5
527	93	23	15	1	20	15	22
528	6	4	3	0	3	2	3
529	39	21	12	4	11	9	12
530	124	25	19	5	14	9	15
531	22	12	0	0	0	0	0
532	34	12	9	4	9	10	8
533	77	21	13	0	8	1	8
534	8	2	0	0	0	0	0
535	46	9	2	1	5	4	4
536	19	11	5	0	2	2	3
537	25	13	1	0	0	0	0
538	2	2	2	0	2	2	2
539	29	7	2	1	3	1	3
540	12	3	1	1	1	0	1
541	372	23	13	9	16	6	15
542	38	1	0	0	0	0	0
543	24	9	0	0	0	0	0
544	324	30	24	21	27	28	26
545	32	11	2	3	5	1	5
546	36	11	2	2	1	2	1
547	144	17	7	2	5	4	4
548	2	2	2	0	2	2	2
549	367	22	9	9	19	17	20
550	60	12	5	0	1	1	3

Table 5: The summary of the official results (526-550).