IRIT at TREC'2002: Web Track

A. Benammar, M. Boughanem, G. Hubert, C. Laffaire, J. Mothe

IRIT-SIG
Campus Univ. Toulouse III
118, Route de Narbonne
F-31062 Toulouse Cedex 4
Email : trec@irit.fr

1 Summary

The tests we performed for TREC'2002 web track focus on the web distillation part. The aim of our participation is to experiment our method for topic distillation combined with a new version of our system Mercure and to validate our system on a large collection of web pages: 18 Go of data.

This year, three runs were submitted to NIST.

2 Mercure model

Mercure is an information retrieval system based on a connexionist approach and modeled by a multi-layered network. The network is composed of a query layer (set of query terms), a term layer (representing the indexing terms) and a document layer [Boughanem99].

Mercure includes the implementation of a retrieval process based on spreading activation forward and backward through the weighted links. Queries and documents can be used either as inputs or outputs. The links between layers are symmetric and their weights are based on *tf-idf* measure inspired from OKAPI [Robertson00] and SMART term weighting.

- the query-term links (at stage s) are weighted as follows:

$$q_{ui}^{(s)} = \begin{cases} \dfrac{nq_u \times qtf_{ui}}{nq_u - qtf_{ui}} & if\,(nq_u > qtf_{ui}) \\ qtf_{ui} & otherwise \end{cases} \tag{1}$$

Where:

- ❑ $q_{ui}^{(s)}$: the weight of the term t_i in the query u at the stage s,
- ❑ qtf_{ui} : the query term frequency of t_i in the query u,
- ❑ nq_u : number of terms in the query u,

- the term-document link weights are expressed by:

$$d_{ij} = \frac{tf_{ij} \times \left(h_1 + h_2 \times \log\left(\dfrac{N}{n_i}\right)\right)}{h_3 + h_4 \times \dfrac{dl_j}{\Delta d} + h_5 \times tf_{ij}} \tag{2}$$

Where:

- ❑ d_{ij}: term-document weight of term t_i and document d_j,

- tf_{ij}: the term frequency of t_i in the document d_j,
- N: the total number of documents,
- n_i: the number of documents containing term t_i,
- h_1, h_2, h_3, h_4 and h_5: constant parameters,
- Δd : average document length.

3 Web Track Experiment

3.1 Indexing methodology

We validate our indexing scripts on the GOV collection. Our scripts have been modified to optimize the build of the dictionary: we obtain now a gain of 50% for the speed-up of the execution.

The queries used for the runs were indexed only with the title field.

3.2 Web methodology

Once the GOV collection has been indexed, three runs were performed and submitted to NIST.

The first one, Mercah, is based on a sample search issued from an evolution of our system Mercure. This run is an ad-hoc retrieval. It has been performed with no relevance feedback and no query expansion

The second one, Mercure, is based on the analysis of the domain of each document in the aim of finding the hit page.

The third one, MercureLynx, was carried out using the results of Mercah ad-hoc search, and then the list of selected documents is re-ranked using the link analysis method we propose. The approach used for this run is described in section 4.

3.3 Web Distillation

Our algorithm for the web distillation experiments is derived from the HITS algorithm proposed by Kleinberg for ranking search engine results [Kleinberg98]. We extend the HITS algorithm to exploit not only links but also the document contents in order to re-rank the document list retrieved by Mercure search engine. The proposed algorithm is composed of the following steps:

1. Neighborhood graph construction: a neighborhood graph is a directed graph consisting of a set of nodes (documents) and directed edges (hypertext links) between nodes. For a given subset of documents, we construct a neighborhood graph containing all the links between documents. The neighborhood is composed only of the documents that appear in the retrieved document set. we consider the 1000 top ranked documents in the first retrieved result.
2. Neighborhood graph analysis: this analysis is based simultaneously on the document links and on the contents. HITS algorithm does not weight the edges of the graph. However, in their experiments, Bharat and Henzinger [Henzinger98] have shown that edge weights improve the precision. Indeed, edge weights reduce the influence of documents that are all contained in one host. In our approach, we define a weighting method that depends on:

- the link typology: In a neighborhood graph, there are two kinds of hypertext links: organizational and navigational. An organizational link relates two documents belonging to the same host (WWW domain) and a navigational link relates two documents belonging to different hosts. In the experiment presented in this paper, we do not consider the organizational links. However, in further experiments, we will consider both organizational and navigational links but giving less importance to organizational ones.
- the link relevance:

The relevance weight of the link from d_1 to d_2 ($wl_{d_1 \rightarrow d_2}$) is calculated as follows:

$$wl_{d_1 \rightarrow d_2} = \beta \times R(d_1, d_2) \qquad (3)$$

Where:
- d_1 and d_2: are documents from the neighborhood graph,
- β: is a parameter used to weight differently organizational and navigational links. β is equal to 1 for navigational links, and to 0 for organizational links.
- $R(d_1, d_2) = Similarity\ \left(\vec{d_1}, \vec{Q}\right) \times Similarity\ \left(\vec{d_2}, \vec{Q}\right)$

We use inner product normalization when evaluating the similarity between a document d_i and the query Q:

$$Similarity\ \left(\vec{d_i}, \vec{Q}\right) = \sum_{j=1}^{t} \left(w_{jq} \times w_{ji}\right) \qquad (4)$$

Where w_{ji} (respectively w_{jq}) corresponds to the *tf-idf* value of the j^{th} term in the document d_i (respectively in the query Q).

4 Results

Table 1 describes the results obtained at TREC'2002 Topic distillation task for officials runs.

Precisions:	*Mercah*	*Mercure*	*MercureLynx*
A 5 documents:	0.2449	0.2041	0.1184
A 10 documents	**0.2163**	**0.1429**	**0.1082**
A 15 documents	0.2041	0.0966	0.0898
A 20 documents	0.1765	0.0724	0.0776
A 30 documents:	0.1463	0.0483	0.0728
A 100 documents	0.0898	0.0145	0.0429
A 200 documents	0.0661	0.0072	0.0293
A 500 documents	0.0356	0.0029	0.0230
A 1000 documents	0.0203	0.0014	0.0203
Exact:	0.1984	0.0575	0.0646

Table 1

It is significant that our best performing run is Mercah which is an ad-hoc retrieval run. We identify two main reasons of these results. First, it confirms that the GOV collection is well indexed and that our scripts are performing. Secondly, it means that best resources for topic distillation such as hit pages can be found with an ad-hoc research engine and are ranked in the top ten documents retrieved. It also confirms that the weight assign to terms is well estimated.

About the run Mercure, it is not surprising that results we obtained are not significant because we retrieved only 415 documents for the 50 queries. Our precision at 5 documents and precision at 10 documents are good because we retrieved less than 10 documents per query. So we can think, we retrieved 10 good resources. Of course, all other precisions are not significant because they are calculated with only 10 documents per topic. It is due to our algorithm which extracts the domain for each document and retrieved only one page per domain. For each query, we found approximately 10 different domains.

The run MercureLynx are not significant. In fact, the proposed algorithm is not perfectly personalized to the topic distillation task. Indeed, the goal of the topic distillation is to retrieve the most relevant resource. However, the aim of our algorithm is to enhance the precision values of the final ranking by including more relevant documents in the first levels of the final result.

However, the experiments have shown the importance of the link analysis precisely by combining the content and hypertext analyses.

5 Conclusion

The results obtained this year in TREC 2002 Web Track show that our system Mercure is able to obtain good results with large collection of data.

It also shows that an ad-hoc research can obtain good results even if it is not specially developed for topic distillation task, so we will work next year to ameliorate our system with evolution such as document structure and query expansion.

References

[Boughanem99] Boughanem M., Chrisment C., Soule-Dupuy C., *Query modification based on relevance back-propagation in ad-hoc environment*, Information Processing and management, 35 (1999), pages 121-139, 1999.

[Henzinger98] Henzinger M., Bharat K., *Improved algorithms for topic distillation in a hyperlinked environment,* 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 101-111, Melbourne, 1998.

[Kleinberg98] Kleinberg J. M., *Authoriatives sources in a hyperlinked environment*, Journal of the ACM, Volume 46, Number 5, pages 604-632, 1998.

[Robertson 00] Robertson S. E., Walker S., *Okapi/Keenbow at TREC-8*, In Proceedings of the TREC-8 Conference, National Institute of Standards and Technology, pages 151-161, 2000.