Mercure at TREC'2003
Web track – Topic Distillation task

MOHAND BOUGHANEM, KAREN SAUVAGNAT, CECILE LAFFAIRE
IRIT- SIG,
118 Rte de Narbonne,
F- 31062 Toulouse CEDEX 4
Email : {bougha, sauvagna, laffaire}@irit.fr

1 – Summary

The tests performed for TREC'2003 web track were focused on the topic distillation part. The aim of our participation is to validate the results we obtained last year and to test the use of term proximity on Mercure model.
As last year, ad-hoc methodologies were used to answer the topic distillation task.
4 runs were submitted to NIST this year.

2 – Mercure model

Mercure is an information retrieval system based on a connexionist approach and modeled by a multi-layer network. The network is composed of a query layer (set of query terms), of a term layer (representing the indexing terms) and of a document layer [Bougha 99].
Mercure includes the implementation of 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 the *tf-idf* measure inspired by OKAPI [Robertson 00] and SMART term weighting.

The query-term links are weighted as follows :

$$q_{ui} = \begin{cases} \frac{nq_u * qtf_{ui}}{nq_u - qtf_{ui}} \text{ if } (\text{nq}_\text{u} > qtf_{ui}) \\ qtf_{ui} \text{ otherwise} \end{cases}$$

Where:
- q_{ui} : the weight of the term t_i in the query u
- qtf_{ui}: the query term frequency of t_i in the query u
- nq_u: the number of terms in the query u

The term-document link weights are expressed by :

$$d_{ij} = \frac{tf_{ij} * (h_1 + h_2 * \log(\frac{N}{n_i}))}{h_3 + h_4 * \frac{dl_j}{\Delta_l} + h_5 * tf_{ij}}$$

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

2.1. Query evaluation

The query evaluation is based on spreading evaluation. Each node computes an input and spreads an output signal. A query is evaluated as follows:

1 – The query u is the input of the network. Each node from the term layer computes an input value from this initial query: $In(t_i) = q_{ui}$ and then an activation value: $Out(t_i) = g(In(t_i))$, where g is the term layer activation function.

2 - Each term node propagates then this activation value to the documents nodes through the term-document links. Each document node computes an input value: $In(d_j) = \sum_i Out(t_i) * d_{ij}$ and an activation value: $Out(d_j) = g(In(d_j))$, where g is the document layer activation function.

The set of retrieved documents, $Output_u$ $(Out(d_1), Out(d_2), \ldots, Out(d_N))$ is then ranked in a decreasing order of the activation value.

2.1. Term proximity

The ranking function (activation) was modified to take into account term proximity in a document [Kean 91]. Thus, documents having query terms close to each others compute a new input value:

$$In(d_j) = \sum_i (Out(t_i) * d_{ij}) * \sum_i \frac{\alpha}{prox_{i,i-1}}$$

Where:
- α is a constant parameter such as $\frac{\alpha}{prox_{i,i-1}} \geq 1$. α is set to 4 for the TREC'2003 experiments.
- $prox_{i,i-1}$ is the number of terms separating the query terms t_i and t_{i-1} in a window of α terms in the document. The query terms are ranked according to their position in the query text.

In other words, documents having close query terms (i.e. no more than α words separate query term t_i and query term t_{i-1} in the document content) increase their input value.

3 – Topic distillation task Experiments

3.1. Indexing methodology

The GOV collection was indexed with scripts allowing to take into account term positions in the documents. Terms are stemmed with Porter algorithm and a stop-word list is used in order to remove non-significant terms from the index.

The queries used for the runs were built from the title and the description fields of the topics.

3.2. Web methodology

4 runs were performed and submitted to NIST. All runs are ad-hoc retrieval, and have been performed with no relevance feedback and no query expansion.
The first two runs are based on a simple search without term proximity. The queries are built using the title field only for the run named *Merc1ti*, and the title and description for the second run named *Merc1td*.
The last two runs, named *Merc2tp* and *Merc2tm*, use term proximity, and only the title field of topics was taking into account.
Merc2tm was performed as follows: if the title field of a topic contains phrases (about 25% of the total number of queries), term proximity is chosen to perform the query, otherwise, simple search without term proximity is used. This choice was made manually.

4 – Results

4.1. Analysis

Table 1 describes the results obtained at TREC'2003 Topic Distillation Task for official runs.

Precision	**Merc1ti**	**Merc1td**	**Merc2tp**	**Merc2tm**
At 5 documents	0.0720	0.0400	0.0800	0.0960
At 10 documents	**0.0720**	**0.0400**	**0.0680**	**0.0760**
At 15 documents	0.0653	0.0360	0.0573	0.0667
At 20 documents	0.0560	0.0330	0.0520	0.0660
At 30 documents	0.0473	0.0313	0.0487	0.0593
At 100 documents	0.0290	0.0216	0.0290	0.0312
At 200 documents	0.0205	0.0145	0.0205	0.0225
At 500 documents	0.0110	0.0088	0.0108	0.0130
At 1000 documents	0.0070	0.0056	0.0069	0.0077
Exact	0.0784	0.0433	0.0669	0.0783

Table 1: Precision at n documents and exact precision for TREC'2003 official runs

Our best performing run for precision at 10 documents is *Merc2tm*, performed using simple search *and* term proximity, depending on the presence of phrases in the query. The R-precision for runs *Merc2tm* and *Merc1ti* is comparable.
The use of term proximity does not affect the results in a significant way.

Using title field only is more suitable than using both title and description fields (run *Merc1td*), as observed last year.
However, it can be noticed that the results we obtained this year are much worse than those obtained in TREC'2002 for the topic distillation task.
Indeed, in TREC'2002 our run *Mercah*, which was an ad-hoc run performed with the Mercure system, was ranked 6th for the precision at 10 documents. The algorithm used for the run *Mercah* is the same as the one used for the run *Merc1ti* performed this year.
The following table compares the precisions obtained in TREC'2002 and those obtained in TREC'2003 for comparable runs.

Precision	**Mercah 2002**	**Merc1ti 2003**	**Merc2tp[1] 2002**	**Merc2tp 2003**	**Merc2tm[1] 2002**	**Merc2tm 2003**
At 5 docs	0.2449	0.0720	0.2122	0.0800	0.2735	0.0960
At 10 docs	0.2163	0.0720	0.1816	0.0680	0.2224	0.0760
At 15 docs	0.2041	0.0653	0.1673	0.0573	0.1878	0.0667
At 20 docs	0.1765	0.0560	0.1378	0.0520	0.1582	0.0660
At 30 docs	0.1463	0.0473	0.1143	0.0487	0.1381	0.0593
At 100 docs	0.0898	0.0290	0.0810	0.0290	0.0882	0.0312
At 200 docs	0.0661	0.0205	/	0.0205	/	0.0225
At 500 docs	0.0356	0.0110	/	0.0108	/	0.0130
At 1000 docs	0.0203	0.0070	0.0185	0.0069	0.0203	0.0077
Exact	**0.1984**	**0.0784**	**0.1542**	**0.0669**	**0.2023**	**0.0783**

Table 2: Comparison of TREC'2002 and TREC'2003 runs

TREC' 2003 precisions are much more lower than those obtained in TREC'2002.

Nevertheless, the analysis of precisions should not be the only criterion of judgement. Table 3 compares the runs *Mercah* (TREC'2002) and *Merc1ti* (TREC'2003) against the published median runs in TREC'2002 and TREC'2003.

	Worst	**< Median**	**Median**	**> Median**	**Best**
2002 (Mercah)	5	2	16	20	6
2003(Merc1ti)	6	0	32	6	6

Table 3: Comparative results at precision at 10

In TREC'2002 and TREC'2003, 6 topics obtain the best results and almost the same number of topics obtain the worst results (5 in TREC'2002 and 6 in TREC'2003).
On the other hand, the distribution between "median" and ">median" is different. Indeed, in TREC'2002 more topics were above the median (20) compared to this year (6). In TREC' 2003, much more topics are exactly on the median.

4.2. Discussion

The first explanation we could give to argue such results comes from this year definition of relevant judgements. Table 4 compares the number of documents considered to be relevant by the assessors for all topics in TREC'2002 and TREC'2003.

[1] These runs were not submitted as official runs in the TREC'2002 web track.

	Number of relevant documents	Average number of relevant documents per topic
2002	1574	32,12
2003	516	10,32

Table 4: Comparison of the number of relevant documents per topic in TREC'2002 and TREC'2003.

In TREC'2003, the average number of documents per topic is 10.32 and has strongly fell down.
For topics with less than 10 documents relevant, the maximum possible scores is less than 1.0. So, it is almost impossible to obtain highest precision this year and it explains our results in term of precision in 2003. So precision at 10 documents is not an appropriate measure.
Let us consider another measure, the R-precision, which has became the main measure in TREC'2003. If we analyze each topic's scores for the best run performed with the original Mercure (*Merc1ti*), we note that for 20% of the topics, the R-precision is relatively high (i.e. higher than 0.2), but for about 60% of the topics, the R-precision is 0. In this last case, most of the topics were those for which the number of relevant documents is smaller than 5. In fact, if there are for example 2 relevant documents, and if they are retrieved at position 3 and 4 by the retrieval system, the R-Precision is 0, even if the relevant documents are in the top 5.

The second explanation of our results could come from the TREC'2003 topic distillation task definition.
Indeed, in [Craswell 02], authors maintain that" *for a few topics, the number of such resources is very much higher than expected. While hand-listed pages in Web directories tended to have short URL' S and high indegree, key resources from this year's track did not show such tendencies as strongly "and in conclusion"[...] the topic distillation task proved difficult to explain to both participants and assessors and there was considerable disparity between the interpretations of these two groups... the task is worth repeating in 2003 but more explanatory effort is needed*".
In TREC' 2002, ad hoc runs could obtain good results on the topic distillation task because the assessors judged more resources relevant than needed and these resources were not always home Page of site or short URL. In TREC'2003, the topic distillation task was better specified [Trec Guidelines 03]: "*We are concentrating solely on websites as resources. The task is to find as many different websites (represented by to their entry pages) as possible within the first ten results.* ".
Thus, considering this definition of the task, ad-hoc runs are no more suitable.

5– Conclusion

The goal of our participation was to test a ranking function based on term proximity on the Mercure model and to validate the results we obtained in TREC'2002 with ad-hoc strategies on the topic distillation task.
The use of term proximity does not improve the results in a significant way.
Moreover, for the 4 ad-hoc runs we submitted this year, performances are not as high as last year. In fact in TREC'2003, the topic distillation task has evolved. Indeed, the spirit of the topic distillation task is to take a large set of relevant results and distill it down to a few key home pages. An overview of the TREC 2003 web track for all participants lets appear that referring anchor text was important and that URL information and link structure was very

useful in several cases. Thus, the top 5 groups at R-precision have used at least document structure, anchor text or link structure. Ad hoc runs show their performances regressed for this topic distillation task.

References

[Bougha 99] : M. BOUGHANEM, C. CHRISMENT, C. SOULE-DUPUY : *Query modification based on relevance back-propagation in ad-hoc environment.* Information Processing and Management, 35 (1999), pages 121-139, 1999.

[Craswell 02]: N. CRASWELL and D. HAWKING : *Overview of the TREC-2002 Web Track*, in Proceedings of TREC'2002. http://trec.nist.gov/pubs/trec11/papers/WEB.OVER.pdf

[Kean 91] : E. Michael KEEN*: The use of term position devices in ranked output experiments*, Journal of Documentation, v.47 n.1, p.1-22, March 1991 .

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

[Trec guidelines 03]: *Guidelines of the TREC'2003 Topic distillation task*. http://es.cmis.csiro.au/TRECWeb/