Criticisms and proposed evolutions for the Information Retrieval evaluation paradigm

Christian Fluhr
Multilingual Multimedia Knowledge Engineering Lab
CEA/LIST CEN Fontenay aux Roses
christian.fluhr@cea.fr

Evaluation / real use

CLEF2005 interrogation : Why as we have good results on crosslingual evaluation, none of the best systems have a commercial success ?

Tentative answer : conditions of test do not reflect the real use of the systems

Evaluation / real use

1 Time between reception of queries and posting of the results authorizes final tuning of the system

Solution : automatic querying by the organizer

2 Time between reception of the data and posting of result authorize a huge amount of work to adapt and tune the system for this data. In real use a very limited amount of time is possible and robust systems that need not adaptation are preferred.

Solution : a class of test were no tuning is possible

Evaluation / real use

3 Response time for query must be less than 3 seconds to be accepted by users

Solution : response time must be taken into account in the evaluation

4 Indexing time is less important but the process must be done in a standard hardware in a reasonable time according to users needs

Determination of relevant documents

Do we agree on the definition of relevant documents

Are they documents that mainly treat the theme ?

Or

are they documents giving an answer even if the document is on an other theme ?

That means : are we looking for documents to inform us about a domain or are we looking for documents containing an answer to our interrogation

Question/answering track can bring a solution

Djoerd Hiemstra : 26/04/99 about TREC8 (Crosslingual track)

Participant	av. Prec. J.	av. Prec. U	uniq rel.
98EITdes	0.1919	0.1962	45
98EITfull	0.2514	0.2767	159
98EITtit	0.1807	0.1841	27
BKYCL7AG	0.2347	0.2406	44
BKYCL7AI	0.2012	0.2184	120
BKYCL7ME	0.3111	0.3391	164
RaliDicAPf2e	0.1405	0.1687	176
TW1E2EF	0.1425	0.1569	107
Ceat7f2	0.1908	0.2319	293
Ibmc17a1	0.2939	0.3168	135
Lan1982	0.0296	0.0487	140
Tno7ddp	0.2174	0.2382	152
Tnoedpx	0.2551	0.2846	109
Umdxeof	0.1448	0.1610	140

Determination of relevant documents

Original systems that found much more unique documents are penalized because we can suppose that their proportion of unjudged relevant documents is in the same proportion that the judged unique.

Solution : judge more documents for the systems that give a huge number of unique one in a first step