SOPHIA in Enterprise Track

Vladimir Dobrynin, Son Pham
St Petersburg State University
35 University avenue,
Petrodvoretz, St Petersburg,198504, Russia
dobrynin@top.apmath.spbu.ru,sonsecure@yahoo.com.sg

David Patterson, Niall Rooney, Mykola Galushka
Northern Ireland Knowledge Engineering Laboratory
University of Ulster, Jordanstown, BT37 OQB, U.K.
{wd.patterson, nf.rooney, mg.galushka}@ulster.ac.uk

SOPHIA participated in TREC 2006 as part of the Enterprise track (Expert search task). Given a topic our task was to find an ordered list of up to 100 experts (from a predefined list of candidate experts) and for every expert create an ordered list of up to 20 support documents. Support document should prove that given person is indeed an expert in the domain presented by the topic.

We implemented 3 algorithms to solve this task which resulted in 3 runs *sophiarun1, sophiarun2* and *sophiarun3*.

All runs are based on Contextual Document Clustering (CDC) algorithm [1,2] applied to a part of W3C document corpus.

Document clustering

W3C collection contains documents of different types. In our experiments we used only two document types: *www* and *lists*. Examples of *www* documents are drafts and final versions of official W3C documents, slides from presentations given by W3C members and so on. Documents of *lists* type are e-mails. We split *www* documents into parts, based on 1000 word long segments and considered every part as a separate document. We didn't split mails (*lists* type documents).

Altogether we have 45,975 www documents before splitting and 122,751 after splitting and 198,394 documents of *lists* type (e-mails).

In parsing we deleted stop-words from a standard stop-list and used Porter stemming. We used the following regular expressions for tokenization: [a-z0-9-']*[a-z][a-z0-9-']*. This means that a term should contain at least one symbol from a-z, can contain symbols from 0-9 and symbol '-'. After stop-words were deleted we replaced the ''' symbol by space and deleted all symbols '-' at the beginning and end of a term.

Altogether we have 495,260 distinct terms.

CDC approach is based on scalable automatic detection of themes presented in a document corpus. A theme is represented by probability distribution of terms that occur in a narrow context. Given a term *x* we define probability distribution of terms that occur in same document with term *x* as its context. We say that this context is narrow if its entropy is relatively small.

In narrow context selection process we set 2,000 as number of narrow contexts (themes) we would like to discover in W3C corpus. We ordered terms by document frequency and skipped 90% of the most rare terms and 0.5% of the most common terms, split all other terms into 7 groups of same size and selected 286 terms with minimum context entropy from every group (giving around 2000 narrow contexts in total). We merged similar narrow contexts (with Jensen-Shannon divergence (JS) between two contexts less than 0.01) producing 938 merged contexts altogether. These merged narrow contexts are considered as cluster attractors. A documents is assigned to a cluster with minimum JS between the cluster attractor and document word probability distribution

Clustering resulted in 779 non-empty clusters with size starting from 1 document (35 clusters) up to 25,648 documents. 716 clusters have size less than 1,000 documents.

Based on these generated clusters and statistics of 2 and 3-words phrases which occurred in their documents, we selected 385,070 "interesting" phrases that relate to each cluster theme.

Our approach to expert search

It is difficult to determine who an expert is, based solely on an analysis of *www* documents. This is because usually all members of a working group are accredited as authors of a document produced by the group but maybe only a few members of the group are actually the real authors. Therefore we used *www* documents as background information to generate narrow contexts, to form clusters, discover interesting phrases and to estimate the relevance of a cluster to a topic. We used *lists* documents only to find experts and to select supporting documents from clusters relevant to the topic.

In our first run *sophiarun1* we used only part of topic – its <title> field.

Given a topic we parsed its <title> field as any other document and then we used inverted index to find all documents (both *www* and *lists* types) that contained at least 2 words from topic title or 1 word if this title contains one word only. We considered such documents as relevant to the topic. Additionally if a document contained the same "interesting" phrase that we had in the topic title then we considered this document as relevant also.

The relevance score of a cluster is calculated as sum of weights of all relevant documents from the cluster. A document's weight is equal to the number of query keywords that occurs in the document plus the number of 2 and 3-word "interesting" phrases from the title that occurs in the document multiplied by 5.

Given the 5 most relevant clusters we get all mails from these clusters that were sent by a candidate expert (we have list of candidate experts with more than 1,000 persons). To calculate a mail score we used 1-JS value where JS is calculated between the mail and topic title word probability distributions. To calculate expert score we used sum of squares of mail scores sent by this expert.

Results sent to TREC is a list of up to 100 experts ordered by expert score (this score should be greater than zero) and for every expert we present list of up to 20 mails sent by this person and ordered by mails scores (mail score should be greater than zero).

In our second run *sophiarun2* we used all parts of the topic including <title>, <description> and <narrative> fields. We don't use here "interesting" phrases. We parsed query as before and calculated JS between the topic and centroid of every cluster. We selected 5 top clusters as relevant and identified a list of experts and support documents as in *sophiarun1*. The only difference is that if for a topic we had a small (less than 20) number of mails (sent by all candidate experts) presented in these 5 top clusters then we try 6^{th}, 7^{th} and so on clusters (up to 20) to get 20 or more such mails.

Our last run *sophiarun3* was a mixture of two previous runs. For every topic we accumulated all mails selected in *sophiarun1* or *sophiarun2* and used them to recalculate scores for all experts. The idea was that for one particular topic the approach implemented in first run may be better then approach implemented in second run while for another topic reverse may be true.

Our Results

Next table presents the results for sophiarun1, sophiarun2 and sophiarun3 in terms of average Precision, R-precision, reciprocal rank, relevant retr @ 5,10, where scores are based on the ranking of candidates without regard to support documents.

Run Id	Map	R-prec	Recip_rank	P@5	P@10
Sophiarun1	0.2248	0.2864	**0.6307**	**0.4980**	**0.4306**
Sophiarun2	**0.1355**	**0.2157**	**0.4674**	**0.3347**	**0.3102**
Sophiarun3	**0.2215**	**0.2842**	**0.6178**	**0.4082**	**0.3980**

Next table presents our results Run1, Run2 and Run3 when candidates retrieved with no positive support documents are considered to be irrelevant..

Run Id	Map	R-prec	Recip_rank	P@5	P@10
Sophiarun1	**0.0934**	**0.1415**	**0.4646**	**0.3184**	**0.2449**
Sophiarun2	**0.0159**	**0.0408**	**0.1398**	**0.0490**	**0.0612**
Sophiarun3	**0.0757**	**0.1314**	**0.3873**	**0.2245**	**0.2061**

Discussion

Presented results show that Run 2 results are worse that Run 1 and Run 3. This means that evaluation of cluster relevance based on number of relevant documents in the

cluster is more efficient than evaluation of relevance based on calculation of Jensen-Shannon divergence between cluster attractor and topic word distribution. The reason may be due to specific style of topic description writing.

In future research we suppose to concentrate on selection of most informative "interesting" words and phrases from topic descriptions and use it as query rather than use row text of topic description.

References

[1] Dobrynin, V., Patterson, D., Rooney, N. Contextual Document Clustering. In *Proceedings of 26th European conference on Information Retrieval Research*, Springer LNCS Vol. 2297, pp. 167-180, 2004.

[2] Rooney, N., Patterson, D., Galushka, M., Dobrynin, V.: A scaleable document clustering approach for large document corpora. Information Processing & Management 42(5): 1163-1175 (2006)