Toward Network Information Navigation Algorithms

Sergei Bel'kov, Sergei Goldstein

Ural Federal University, Yekaterinburg, Russia,
srgb@mail.ru, vtsl@dpt.ustu.ru

Abstract. Attention is paid to the problems of automatic search of documents by search engines, analysis of documents and the use and developing of network resources, such as thesauruses and ontologies. Also some of proposals to expand the conceptual model associated with the need to reduce the dimension the set of documents found by the search engine to a set of relevant documents are formed.

Keywords: search engines, query optimization, analysis of documents.

1 Introduction

The numbers of ways we use the Internet now-a-days are really extensive. However, algorithms associated with that are almost not formalized and therefore there are many unresolved problems here. Therefore, the possibility of improving this situation, it is important.

In complex cases we have rather complicated query, and the output is the set of retrieved documents, many of which could not viewed physically, or they have duplicates of other documents or not useful for our tasks.

2 Problem of informational navigation

Problems which associated with informational navigation include there are three main components:

- Search of information, i.e. some documents or texts (books, journals, proceedings, web resources, search methods);
- Analysis of information (formats for documents, methods of obtaining the set of relevant texts, analysis methods);
- The work with network resources (dictionaries and reference books, standalone and web thesauri or ontologies).

Traditional search process (SP) of documents on the Internet can be presented by three components:

$$SP =< Q, SE, DOC >, \tag{1}$$

where Q - the set of queries; SE - many search engines; DOC - found resulting links to documents (further documents).

Query q usually includes a list of simple keywords or phrases which made up by the disjunction of conjuncts or disjunctive normal form.

Search methods which hiding within known specific search engines usually are not obvious to the user.

In addition the found resulting documents can be presented in different formats (txt, doc, pdf, ps, djvu, html, xml and others). Also a set of documents obtained by different search engines in response to the same query may vary essentially.

This raises the following tasks: selection of the most effective (in terms of search target) search engine; optimization of the structure of the query; selection from the set of received documents to only those documents that best meet the targets of the search.

Tasks associated with the optimization of the structure of the query and reducing of the set of received documents are usually beyond the capability search engines.

To resolve some of those problems we suggest to introduce feedbacks into the traditional search scheme (Fig. 1).

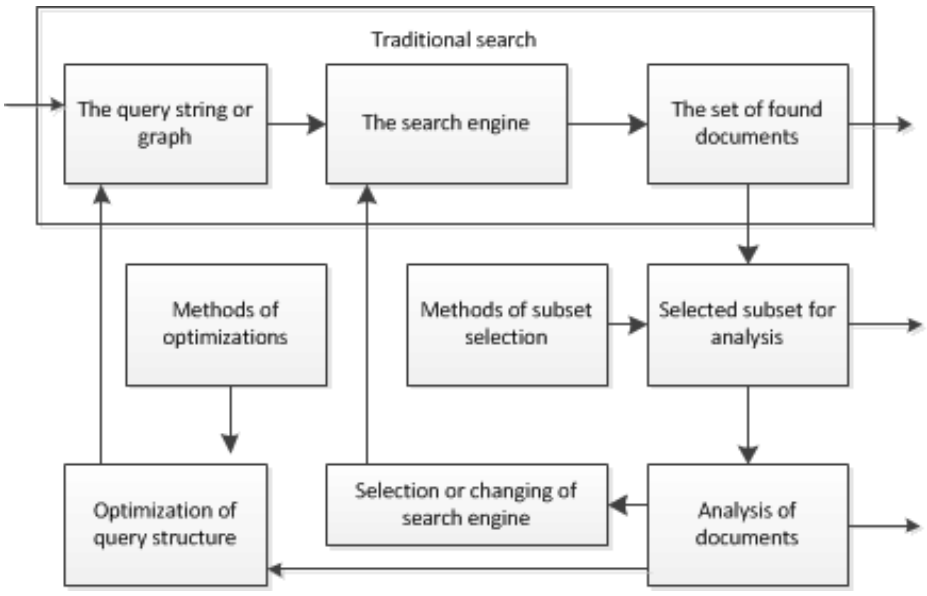

Fig. 1. Search scheme with feedbacks (the top three units are traditional)

With this in mind the procession model takes the following form:

$$SP =< Q, SE, DOC, DSM, DS, MA, A, SES, MO, QO, RDOC >, \quad (2)$$

where first three traditional components was given above; DSM is methods of subset selection for documents analysis ; DS is selection procedure of documents; MA - methods of analysis into selected subset; A is analysis procedure ; SES - selection or changing of the search engine; MO - optimization methods of query structure; QO - optimization procedure; RDOC is the resulted set of relevant documents.

Consider also some of these components separately we may suggest also some useful formalisms.

Serious problem for the analysis may be large dimension of the set of documents on the search engine output. Restricting the sample may be analyzed by random selection of documents involving experts or require the development of additional procedures.

To select a specific search engine, may write:

$$SS_k = F_{sel}(SE, C_{sel}), \tag{3}$$

where F_{sel} - select function ; SE - many of available search engines; C_{sel} - selection criteria.

The result of analysis of the set of documents obtained by applying the k-th search engine:

$$R_k = F_a(DOC_k, C_a, M_a), \tag{4}$$

where F_a - function analysis ; DOC_k - set of received documents; C_a - criteria of analysis; M_a - methods of analysis.

We also introduce the concept of optimal query:

$$Q_{opt} = F_{opt}(R_k), \tag{5}$$

where F_{opt} - optimization function of the query structure (for example graph of connections between keywords).

Often set of the found documents DOC_k is too large (typically tens of thousands). Therefore, one of the optimality criteria is to reduce the number of documents which obtained by query. Other criteria can be adequacy to the search target and complete-ness of the topic consideration.

With this in mind we may get the Algorithm of informational (text) search (Fig. 3). It is algorithm of first-level of decomposition.

At first it may demand some studies of search models or search query languages.

After that we may use for example one of the following search models: search by keys, wide primary search, random wide primary search, intellectual search, search by last heuristic, search by random walks and others types of search.

Obtained results may be divided into several groups depending on the different cri-teria or search characteristics.

Working with set of found documents will demand methods of documental analysis. During the analysis of the set of documents may appear the following tasks:

– To identify the documents which are most similar to the search aims. That may be such documents as at random taking a number of documents from the beginning of the set (for some search engines, they are usually the most relevant purpose of the request). Also more special procedures may use here (for example by taking documents one of presentation format);
– Divide the set of documents for the group (for example: unimportant, secondary importance, and high importance documents), areas or classes.

It uses a set of keywords or phrases (terms), which are presented in the documents. Some of these terms are also present in the query q. Document is describing its set of keywords is the image of the document.

For domain we have a Dictionary, consisting of terms. To determine the degree of connection between the two documents apply the mathematical apparatus of the following models: Boolean, Extended Boolean, Vectoral, Fuzzy logical, Probabilistic [1]. Nevertheless, a direct comparison of these methods is difficult, it requires the development of additional mathematical apparatus. In more complex cases, the dictionary is transformed into thesaurus or ontology. For hypertext some special form patterns may used [2].

The resulting images of documents are allowed to move to the problem of classification. There are images of reference documents (supervised learning) or clustering of documents where no master images (learning without a teacher).

The resulting matrix of pairwise proximity of documents allow us to go to their classification or clustering. Thus we gave the following tasks: exclusion of non-uninformative (in terms of search target) documents (infor-mation noise); elimination of duplicate documents; partition (classification) of the set of documents into two (important, unim-portant) or three main categories (low, medium and high degree of importance); the actual clustering as a partition of the set of documents into groups according to the properties of their images (feature vectors).

Many of the docu-ments can be excluded on the basis of viewing only its Title or Abstract. Thus there are presented three levels of consideration: primary, main and additional analysis. Turning to the image of the document as a set of keywords, we also have several levels of keywords analysis: top (from the query), middle (from Abstract) and low (from text content, i.e. known and new keywords).

Thus after analyzing the problems arising from modern network navigation we proposed to complement existing search engines several of additional units, in particular helping to optimize the structure of the query and limit the set of relevant documents.

We plan to consider them in detail in our further studies.

References

1. *Lande, D. V., Snarskii, A. A., Bezsudnov, I. V.:* Internetika: navigation in complex networks. Librokom, Moscow (2009) (in Russian).
2. *Belkov, S. A., Goldstein, S. L.:* Representation of materials of text and hypertext sources by net of patterns. J. Informational Technologies. 1 (161), p.29-34 (2010).

Алгоритмы сетевой информационной навигации

Сергей Бельков, Сергей Гольдштейн

Уральский федеральный университет, Екатеринбург, Россия,
srgb@mail.ru, vtsl@dpt.ustu.ru

Аннотация В работе представлен перечень основных компонентов информационной навигации в сети Internet. Рассмотрены вопросы оптимизации поисковых запросов и самого процесса поиска. Представлена расширенная кортежная модель поиска. Предложено несколько полезных формализмов.

Ключевые слова: поисковые машины, оптимизация запроса, анализ документов.