Intelligent search based on ontological resources and graph models

Chugunov A.P.
Computer Science Department
Perm State National Research University
Perm, Russia
chugunov@permedu.ru

Lanin V.V.
Department of Business Informatics
National Research University Higher School of Economics
Perm, Russia
lanin@perm.ru

***Abstract*— This paper describes our approach to document search based on the ontological resources and graph models. The approach is applicable in local networks and local computers. It can be useful for ontology engineering specialists or search specialists.**

Keywords—ontology; semantic; search; graph; document.

I. INTRODUCTION

Today the amount of electronic documents is very large and information searching remains to be a very hard problem. The majority of search algorithms, applicable in local networks, based on full-text search and don't take into account the semantics of a query or document. And good statistical methods can't be used in the local documents repository.

Mathematical and statistical (latent semantic search), graph (the set of documents presented as directed graph), ontological (the searching by existing ontologies) methods are used in computer search [1]. All of them have some imperfection [2].

In spite of this, the tandem of latent semantic and graph methods give very good results. The majority of internet search engines use it [3]. But graph method is not applicable in local networks or local computers [2]. And this approach doesn't let the consideration of semantic context of documents or all parts of search.

So, the task of semantic search hasn't been decided yet. And the newest search algorithms on the internet remain inapplicable in local networks or local computers.

If we combine the tandem with the third, semantic method, we get a possibility to decide the problem of taking into account a semantics. We have chosen ontologies as a semantic method because it allows solving the problem of a document directed graph building too. The building of full ontologies is not required.

The aim of our survey is to unite three different search approaches into one.

II. DESCRIPTION OF RELATED WORK

We observed the most popular algorithms of different search approaches:

1. Namestnikov's A. M. algorithm informational search in semantic project repository [4];
2. information search based on semantic metadescription [5];
3. In-Degree algorithm [6];
4. PageRank algorithm [6];
5. HITS algorithm [7].

The survey was made with a tendency towards on ontology applicable in approach, precision and recall of search results. The extract of survey [3] is presented in table 1.

TABLE I. THE SURVEY OF SEARCH ALGORITHMS.

	Using of ontologies	**Ontology applicable**	**Precision**	**Recall**
Namestnikov's A. M. algorithm informational search in semantic project repository	Yes	Yes	85%	69%
Information search based on semantic metadescription	Yes	Yes	97%	85%
In-Degree algorithm	No	Yes	75%	47%
PageRank algorithm	No	Yes	81%	66%
HITS algorithm	No	Yes	63%	78%

The highest result of precision made by information search based on semantic metadescription. But this algorithm requires a lot of ontology building, because it needs human participation. [2]

So, we decided to use HITS algorithm, because it has the best result and it's applicable to our work.

The using of ontologies in HITS algorithm is planned on the stage of primary documents set forming, which satisfy the query, as well as on the stage of G_δ forming and changing.

III. Definition

We used the following definition of ontology [5] as basic: ontology is a triplet $O=<X,R,F>$ where

X – not empty set of concepts of subject area;

R – finite set of relations between concepts;

F – finite set of interpretation functions, adjusted on concepts and/or relations of ontology;

We must mention the fact, that R and F can be empty. Ontology can contain instances of classes – the classes with preset properties.

In our work we will use the changed definition of ontology: ontology is a pair $O=<X,R>$ with some constraints on the concepts set and relations set [2].

Document in our paper is a set of properties of a real document, subject, content and document ontology. Properties of real document are any data about it, which isn't presented in the content, including metadata.

$$D=<R,C,O>$$

R – set of document properties. A set of properties can be described by metadata standard "Dublin core" [4];

C – content, i.e. entry of the document;

O – document ontology.

IV. Algorithm describing

Proposed algorithm consists of 5 steps:

1. Building ontology O by existing documents.
2. The second step is to enter a query by the user, i.e. the determination of the set of primary concepts $\{C_i\}$, is interesting for the user.
3. The third step is the allocation of a documents set A_i, that contains all or some from $\{C_i\}$. Denote this A_s.
4. The fourth step is executing a range algorithm with input: $\{C_i\}$ as a user query, A_s as a primary document set, O as a directed document graph.
5. Output results to the user.

A. *Building ontology*

The step is preparatory. On this step we solve the task of automatic document ontology building, selected document properties from unstructured text on the natural language.

After document ontology building we determine "link to" type links between documents. It is advisable to combine this links into separate ontology. During this process not existing files can be included in the set. These links must be placed in the set because it allows making search of documents, which doesn't exist in the repository.

After that we get 2 levels of ontologies:

1. Document ontologies. Define them $\{O_1, O_2, \dots O_n\}$, where n is amount of documents.
2. Documents links ontology. Define it $O_{L.}$

In addition, the subject area ontology O_p can be made. This ontology doesn't depend on documents D, it contains only the knowledge about the subject area. Building of O_p can be automatic or manual. It's main, that if the amount of documents subject areas will be large, large amount of O_p can spoil the results. It leads to anomalies, conflicts, ambiguity between ontologies.

B. *Entry a query by user*

It's the first step in search. The aim of that is to determine a set of concepts $Q=\{C_i\}$, which are interesting for the user.

From the query we select keywords, concepts. Next, we extend the set due to subject ontologies, if it exists. This extend will contain synonyms, definitions and else.

Besides, the part of ontology O_p is being built on this step. The part contains a user query. Afterwards we will use it for calculating the weight of documents.

C. *Allocation of a documents set*

The goal of this step is building a primary documents set $D_F=\{D_i\}$, which satisfies the user query Q. The set is not final and can be changed on the next step.

Since the set is not final, we use latent semantic search on this step. We choose it because it gives high speed of search and relatively high precision of results.

The primary set can be calculated by the following formula:

$$D_F = \{D_i | X \cap Q \neq \emptyset, X \epsilon O, O \subset D_i\}$$

In this set come documents, which keywords and concepts X in document ontology O (we define it O_i) are crossing with $\{C_i\}$ in the user query Q.

After that, we assign weight of each document in D_F. This weight reflects semantic distance to user query. This weight can be calculated by

$$w_i = avg(sim_{sem}(t_1, t_2) | \, t_1 \in O_i \text{и} \, t_2 \in Q),$$

where

$$sim_{sem}(t_1, t_2) = \begin{cases} |k| \cdot \dfrac{sim_{sem}(s_1, s_2) + sim_{sem}(o_1, o_2)}{2}, if\ k > 0 \\ |k| \cdot \dfrac{sim_{sem}(s_1, o_2) + sim_{sem}(o_1, s_2)}{2}, else \end{cases}$$

where $k = sim_{sem}(p_1,p_2)$ is value of distance between predicators, and t_1, t_2 are triplets. Triplet is a set of three $<X_1, P, X_2>$ where X_1 and X_2 are ontology concepts, P is predicate, relation between X_1 and $X_{2.}$

User query Q and documents D_i have ontology view O_Q and O_i. Each ontology divides into triplets t_1 and t_2, which can be intersected in an ontology. Next, we calculate semantic distance in pairs. Semantic distance between a user query and a document calculation as average of semantic distances of them triplets. It allows to take into account not absolute coincidences.

If we combine O_L and $\{w_i\}$, we get weighted directed graph $G=<V,E>$, where V is a set of documents $\{D_i\}$, some of them has a weight – a number. If number is missing, the weight we let 0. Set E – a set of directed arcs, which present the links between documents. Arcs E haven't weights because it's impossible to determine power of link automatically with needed accuracy between documents today.

D. *Executing range algorithm*

Primary documents set D_F are extending by documents, which have links (in or out) with documents from D_F. In algorithm exists parameter d – amount of documents, which can be added by document from R_δ. In the set must be added d or fewer documents with maximal weights (semantic distance). It's important, that the weight of adding document must be bigger than w_{min}. This rule rises precision and recall of the results.

Documents ranging process base on vertex weights and amount of in- and out- arcs. It allows to get semantically closer documents in the results, even if they have small amount of arcs or haven't them at al.

So, the result of the algorithm is a set of pairs $D_R=<D_i,r_i>$, where

D_i – found document

r_i – rang of the document.

E. *Output results to the user*

The set D_R can be output to the user as a traditional list of documents ordered by their weights or in graphical mode – as a document graph.

V. Conclusion

In this work we developed, offered and described our information and documents search approach, which combine 3 most widespread methods. We described it mathematically.

Now we have started the first realization of this approach. As a starting subject area we have chosen the science papers and publications, because these documents meet the standards of typography.

References

1. Gasanov E. E. Information storage and search complexity theory, Fundamentalnaya i prikladnaya matematika, vol. 15 (2009), no. 3, pp. 49–73.
2. Никоненко А.А. Обзор баз знаний онтологического типа / Штучний інтелект, 2009, № 4. С. 208-219.
3. Signorini A. A survey of Ranking Algorithms, http://homepage.divms.uiowa.edu/~asignori/phd/report/a-survey-of-ranking-algorithms.pdf
4. Наместников А. М. Интеллектуальный сетевой архив электронных информационных ресурсов / А. М. Наместников, Н. В. Корунова, А. В. Чекина // Программные продукты и системы, 2007, № 4. С. 10-13.
5. Гладун А.Я. Онтологии в корпоративных системах / А.Я. Гладун, Ю.В. Рогушина // Корпоративные системы. – 2006. – № 1. – С. 41-47.
6. K. Bharat, Henzinger M. R. Improved algorithms for topic distillation in a hyperlinked environment// In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, 1998, New York, USA, p. 104-111.
7. Kleinberg J. Authoritative sources in a hyperlinked environment, Journal of ACM (JASM), №4, 1999, pp. 85-86.