The Georgian Semantic Search Engine Development – Complexity and Decision

Manana Khachidze[1], Magda Tsintsadze[2], Maia Archuadze[3], Gela Besiashvili[4]

Department of Computer Sciences, Iv.Javakhishvili Tbilisi State University, Georgia

[1]manana.khachidze@tsu.ge, [2]magda.tsintsadze@tsu.ge, [3] maia.archuadze@tsu.ge, [4]gela.besiashvili@tsu.ge

***Abstract* – In the proposed article the difficulties of Georgian language based semantic retrieval "engine" algorithm development issues are being considered. The detailed research work plan on Georgian Language web based Corp development/update system for general using purposes is provided. The Corp will be serving as general source for Georgian language based search engines. It is offered to develop modified labeling for non-structured documents using method of Analytical Heuristics in order to form the concept "patterns" knowledge-base. This method of indexing will allow presentation of non-SWD documents in form of "pseudo-SWD" documents and little external engine on SWSS will give an opportunity to develop the semantic retrieval algorithm based on concept patterns categorized by various subject ontologies.**

We suppose the best online retrieval system in future will be the one oriented on semantic web, thus the natural language Corp based semantic search systems are of uncontested perspective.

Index Terms: information retrieval, semantic search, semantic web,

I. INTODUCTION

Information Retrieval represents the classical problem of Informatics. In accordance of giant information flow, the importance of semantic search rises as well. Despite the fact that research in this field is performed for quite a long time and actively, the search engines are not still perfect, thus the actuality of search engine improvement and new algorith processing is a quite modern task aspecially for those systems that are directly involved in the process and/or performing it firstheand.

Generally speaking all main standard text retrieval systems consists of three different files: first is the full text of document along with full bibliographical and indexed information, second is the dictionary of all unique words of text sorted alphabetically and the last file is the inverted list that preserves the position of each word in alphabet. This structure is called inverted and the search performed using inverted data structure (especially in large documents) is of highest importance as each word might serve as the start point in retrieval process.

II. IMPORTANCE OF THE PROBLEM

Search engines are using document key word base query searching, in fact this method means to define string (lexical) conformity between search query and internet document included terms. Information retrieval based on key-words are used for non structured web-documents in general.

One of the most popular internet-searching method is the Bool retrieval, that is based on key-word different combination discovery using AND, OR, NOT operations. There are also many other types of searching tools like: Wildcard Symbol, statistics based methods (Google ranking), context retrieval, saga-oriented searching, searching based on key-word position and etc.

Information retrieval based on above mentioned techniques require indexation of billions (and the number is growing each second) of web-pages, thus the web based document content analyzing problem is getting more and more important. To dig a little deeper in information retrieval history we will find out that the importance of internet-document content analysis rises in accordance of huge increase of web-page number. In 2001 the semantic web technology was presented by W3C (World Wide Web Consortium), according to their definition the main target of these technologies is to transform non-structured or semi-structured web-documents into the "web of data". Semantic web is based on Resource Description Framework (RDF) that represents the standard model of information exchange; it has peculiarity to unite different schemas based data. RDF is generalizing link based structure of the document and is using URI for relationship and metadata for websites, it allows representation and sharing of structured and semi-structured data among different applications.

The main purpose of semantic web is to generalize existed www by inserting metadata and making easier for machines (search engines or any other automatic agents) to "understand" and respond accordingly to human requests. To accomplish this objective and "understand" the human requested query, the appropriate information source should be semantically structured. Thus using tools specially designed for data like Resource Description Framework (RDF), Web Ontology Language (OWL), and Extensible Markup Language (XML) it is possible to describe things like animals, cars, people and even parts of the construction in contrast to HTML, that is describing documents and links between them. The terms metadata, ontologies, semantics and semantic web are inseparable.

The semantic web search technologies (Onto Search, Semantic Portals, Semantic Wikis, Multiagent P2P) proposed by the experts of semantic retrieval are using ontologies to form the knowledge base. The semantic web document might be defined as a document that has ontology as its content or as an ordinary web document "labeled" with specific tags taken from the so called Domain Ontology. According to ontology types used for labeling (make annotations) the annotated web documents can be divided into different types.

The existed semantic technologies are trying to solve two problems: formal query authomatic formation (query-ontology) and dealing with such collection of web-documents, the structure of which is not forehand defined.

Keyword-based search technology might also be used in SWD retrieval by matching query terms to terms that lexicalise ontology elements in a document. Swoogle search engine is using this method of searching where the semantics of the SWD is not used, instead the term similarity is defined using the lexical methods, but for the semantic search algorithm term lexical and syntactic similarity is less important, main role here plays their "meaning" similarity.

Document's semantical concept knowledge is necessary for semantic matching. In case of formal query each term semantics might be defined in an explicit way, thus if we have to deal with ontology–query, then each term concept is defined by semantic relation between this term and other term ontologies. Such a relation is not only „is-a", but also „part-of", „meronym", „synonym" end etc. In case of informal query, for example query based on natural language, each query containing term semantics should be somehow defined. The problem is how machine will handle problem of query "real meaning" understanding to provide request-appropriate document.

For semantic reference (on additional request) document should also contain its own semantics. In case of Semantic web document the semantics are defined in ontology in formal and disambiguate way. It is necessary to use modern ontology technologies for non-structured documents.

Nowadays the internet search generalized algorithm for SWSS (Semantic Web Search System) system is developed. It unites several technologies and allows us to create a meta-engine for semantic web document filtration received by SWOOGLE search system.

SWOOGLE is a crawler-based indexing and retrieval system for SWDs in RDF(S), DAML, or OWL syntax. It provides techniques of documents semantic referencing prior the query execution.

As all authomatic systems, the retrieval systems are also facing its Achiles heel as they have to deal with natural language with its appropreate specifications and "whims". Problems posed by natural language are very hard and almost impossible for solution without getting the real meaning of the words in a search query. Uncertainty is becaming more crutial when the system fails to gain knowledge and can't perform real word "human way" perseption. Semantic search systems can solve some of most common problems associating with information retrieval:

- Too much Synonyms

All that we like so much in natural language is representing a big problem in information retrieval, for instance according to writers skills same meaning might be represented in several variations of text, thus search engine is facing problem of understanding what user was really meaning while performing this or that query search . Also the existing synonyms of the same word for different professionals of different countries may vary. All Natural languages have this problem and Georgian especially as for different corner of Georgia we have different words for the same meaning, so we are in need of retrieval system that is able to catch the idea and not the word itself. The semantic search system should handle this problem by generalization of the searched query with synonyms.

- Polisemy

  In Georgian Language as well as in all natural languages one word might have several meaning according to the context it presents in. Semantic search systems are able to perform query compilation in accordance of its context and may serve as tool for solving polysemy.

  Despite the fact that Semantic web provides better solution for information retrieval, we have to mention some challenges that it poses and even this minute the active research/ work by WC3 is in process to handle them :

- The already existed huge amount of unstructured internet-documents requiring the labeling in order to use the semantic search system. Partially this problem is solved as already there have been presented automatic "labeling" systems that can on fly transform structured query to RDF.
- Free form processed query (request provided on natural language or the set of key-words) authomatic transformation method development and etc.

In the process of semantic search system development and release some additional problems appearing:

1. Useage of external resources. SWSS should cover additional external resources if the query performed in natural language is used for it. This kind of knowledge base might be presented in form of general vocabulary and/or the referenced ontology.
2. Automatization and transparency. SWSS has to provide whole process transparently from the beginning up to an end via deriving the ranking list of semantic search documents that match each query. Semantic web document searching process has to be

performed with minimum human participation: user should only validate the output data.

3. Performance. The process of semantic search should be quite fast: real-time querying is of a high importance. SWSS has to be realized as malty-step process with Retrieval of SWDs must be performed with short processing time to obtain the optimal result.
4. Precision. SWSS system accuracy is important as well. Semantic web document querying in real systems, used technologies and realization testing, along with implementation should be performed in respect to precision/recall. Particularly the automatic retrieval of semantic search documents and automatic disambiguation of terms should grant the high precision of results.

In Georgian language performed retrieval problems should be stated separately. Simple experiment using Google search will show that same information retrieval, based on Georgian language query and English language query will derive quite different results. The reasons are several: very low Georgian language web documents presented in semantic web document form, the morphological-syntactic complexity of Georgian language, no or very little SEO and least but not last very important problem of Georgian language corps: there are couple Georgian Language corps, but they are not available online, and can't be used by search engines to form dictionaries and ontologies not only for retrieval purposes, but for automatic translation also.

We suppose in retrieval processes without semantics human went as far as we are now: databases are getting larger and larger, the search engeenes returened documents of bigger and bigger sizes. Despite the fact that Bool retrieval method allowing us to shrink the information to real time processing size, great part of the loosed information might potentially be of a high importance, from another hand statistical retrieval methods based outcome, even they are ranged, are huge. The semantic search is the new tool to interpret query in such a way that represents the full content and puts in correspondence the appropriate knowledgebase.

The semantic search based systems are able to improve statistical search and make results more precise, in the same time they have potential to generalize the Bool search and derive better outcome. Our retrieval algorithm based on natural language Corp will play the significant role in search result improvement as it will provide document and/or query semantic interpretation.

From our point of view the best search engeene in future would be the one that can combine different technologies with semantic web oriented features. We think that natural language Corp base semantic search system is the best solution so far.

III. The object of research for engine development

The main target of the proposed research is the information retrieval, particularly Georgian language processed search system "engeene" algorithm development. The research strategy is based on better development potential technology, realization of which is possible in frames of semantic web and the received results are oriented on recent(near future) challenges.

The stages of reserch relize and tasks are as follows:

1. Semantic web based information retrieval modern technology analysis. Studying and analysis of last relized (last date:2014-02-25) RDF, RDF Schema 1.1 standards and their referenced issues
2. Georgian Language Corp development
3. The text(document) labeling (meta-data) method development
4. The semantic search algorithm development
5. Result Analysis and software realization.

Let us define each task separately:

A. Semantic Web Based Information Retrieval Modern Technology Analysis

Keeping in mind that semantic web was developed as an extention of ordinary web and it is oriented on data (not document) web, the information search process might be presented as the sequence of the following stages:

- ✓ The required information analysis and appropriate query formation
- ✓ The definition of information array source(s)
- ✓ The information selection process from defined arrays
- ✓ The retrieved information and retrieval result analysis

We have to note that the recall of the process is fully dependt on each level success. As we mentioned earlier semantic web is based on specialy developed framework RDF. Thus the search process realization based on this technologie requires the modern standard relize studying.

B. Georgian Language Corp Development

As the main target of the algorithm development process is the georgian language processeed search system "engeen" development available online, the significant part of the research has to be devoted to Georgian language corp elaboration. Sometimes it is asumed that the Web itself might be used as a language corp. We count on creation of such "engine" that enables us automatically collect large size data (text, document), "label" it and make their linguistic parameterization. The appointed issue requires the following stages to be fulfilled:

1. appointing the word-form in searched text;
2. appropriate lexeme defining;
3. Morphological parameter selection;
4. The ability to fix the position in respect to other lexeme;
5. The ability to fix the position in a sentence;
6. Definition of number of sentence/paragraph in text
7. The text creation data definition

To achieve the above mentioned goals the following work should be implemented:

1. The Georgian language database development based on web searched texts and their classification using the general parameters. Finally it will be transformed to web document Georgian language corp. Alternatively it is possible to create a model of the data bank based on the quantum database.
2. Staduing and analysis of common and recent method of indexation-labeling, their pros and cons along with their paculatiries.
3. The analytical heuristics method based labeling algorithm formation and analysis (The georgian language database created in frames of this investigation will be used in an algorithm processing) .
4. The algorithm testing: The testing should be confiremed on two tipes of databases: 1) On the georgian language database created in frames of this research and 2) On general web documents
5. The existed semantic search algorithm stadying and analysis.
6. The conceptual sceme development for ontologies. The knowledge subject area formalization, data structuring method development using concept-pattern formation algorithm. The quantum concepts might also be used in this process in order to create a complete mechanism of concept formation to formalize the sector areas of knowledge.
7. Ontotlogy knowledge base based semantic search algorith development
8. The semantic search algorithm testing fulfilled in two directions: 1) For Georgian language based texts ; 2) Web documents
9. The development of online Georgien language corp – web portal need to be created, it will give possibility to use the georgian language corp for georgian language processed information retrieval purposes.

C. Text labeling method development (metadata) – It means development of such indexes or document defining metadata structures, that might be used for search "engines". Thus we will need to analyze the following factors of common/recent indexation/labeling:

- The forward index analysis : as the forward index stores the words for each document, the process of document parsing should be analyzed. It is important as different web-pages are of a different structure and non-structured documents are hardly defined in relevant result retrieved list.
- The compression method analysis is important as well. According to compression techniques the label might be reduced significantly. In case of large scale searching and huge databases comparison plays quite important role as in fact it is a measure of cost.
- The releability issues should be analized.

In the same time the alalysis of different type lebeling data structure, especialy when we have to deal with such lexicaly-difficult language as Georgian, is required. Modern search engeens are using the following meta-data structures: inverted indexes, sufix-trees, citing index,N-grama structure, document term matrix and etc.

Our offered method of text labeling is based on text semantic analysis and its representation with help of concept-patterns. The "pattern" formation method is analizing the natural language full text and is creating structure containing diverse words, that describing the concept in generalized form. The alanizing process held for different type meta-data structure will allow us to refer the lebeld data structure to one of already existed types or represent it as a new hybrid-model.

D. The Sementic Search Algorithm Devepment

The semantic search algorithm development by recent semantic search method analysis and generalization using our proposed method is the second part of our proposal. The main peculiarity of the semantic search is the document conceptual representation that is formed using knowledge semantic models of the appropriate subject area. The ontology represents the most common and convenient tool among the knowledgebase presentation existed toolkit. Generally subject area defining knowledge is described by concept and property hierarchy, but the joined concept instance semantic net might be used as well. The ontology techniques might better the search quality that is why it is so important to develop the knowledge base subject area defining algorithm In our project the proposed algorithm is based on the analytic heuristics method. [6].

E. The Result Analysis and Software Application

Sure in a real case we will have much more words in a concept description, but the selection of all high weight words is possible using the Explicit Semantic Analysis (ESA) method. The number of words might also depend on ratio of text word number to different word number in the same text. More words presented in concept description will lead to more semantically adequate results, but from other hand many word in description might lead to case when concept will be useless for information retrieval.

In order to define the optimal number of distinct words in concept definition and analyze the results the software realization is necessary.

To evaluate the method its testing should be performed and the testing stages has to be as follows: 1. Concept Formation; 2. Retrival according to formed concept.

As the concept represents the implicant (disjunction, conjunction) Bool algorithm might be applied, thus this is the method we are offering to use for method testing.

IV. Expected outcomes of the research

Keeping in mind that there is almost no Georgian language based search systems and diverse Internet search engines are providing non favorable results, the value of the proposed research seems quite impressive.

The proposed search “engine” application will be possible not olny for web search engines, but it would be convinient for Georgian language based web-sites, where the relevant information retrieval is needed.

As the main work principle of the search engines is based on query-document comparison, the result, gained by author of the query is covering only those documents that are containing one or more words of the searched query. Results based on such method of retrieval are not precise and might contain lot more documents that are out of the users interest and vice versa, might not cover all those documents that in fact are in the interest area of the user. For the best result the definition of contextual connection and reference is necessary. With help of proposed algorithm the concept defining database according to subject area will be constructed. Then the developed ‘engine’ will allow to use not anly the search query contained words, but their appropreate concept pattern elements aswell. The engine will automaticaly modify(generalize) the user search query, thus user will have opportunity to receive all the documents even those ones not containing the search query words directly but are appropleate to search request in a contextual meaning.

The received method application will be possible for text fragment structuring and analysis. It requires the text arrays to be splitted (might be considered the hierarcy structure) into subsets by some contextual selection (the automatic clasterization). We are counting on analytical heuristics method to fulfill this task. The application of this method for the texts will lead to the term vector presented in more compact form comparing to the vector received from ordinary proximity metrics application.

The developed algorithm might be quite useful for other type research and applications. One of such directions might be named the text latent-semantic analysis.

The latent semantic analysis is important not only for linguistic research, but the marketing research aswell. Recently the so called questionaries (survey) containing evaluation questions about the product are in use for new product stadyings. In such a type surveys respodents are evaluating the product not using the pre defined parameters and appropreate points, but sharing their point of view on usefulness of this product different features. The documents received on the basis of such texts are directly connected with text latent semantic analysis.

The latent-semantic analysis of the text is important for different expert systems and decision making information systems in process of knowledge base formation. In several subject areas the knowledge selection is harder because of their low formaliasion possibilities. The expert astimations in common case represents the text, that requires the latent-analysis in order to select the appropreate knowledge. The concept pattern formation algorithm might be used to formalize this knowledge and represent it as a knowledgebase. It might be performed in the same way we suggested to fill the ontology base for the search system.

Finally, in frames of the research the Georgiean Language Corp will be developed, that will be mainly filled with Georgian language texts from the www. If we underline ones more the fact that there is no Georgian Language National Corp and that it is impossible to perform the language research process without its concrete performance statistical data analysis, this part of the research should also be couted as quite important.

References

[1]. Manning, C. D.; Raghavan, P.; Schutze, H. Introduction to Information Retrieval, Cambridge University Press. 2008.

[2]. Cavnar W.B., Trenkle J.M. N-Gram-Based Text Categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994

[3]. V.V. Chavchanidze, (1974) "Towards The General Theory Of Conceptual Systems: (A New Point of View)", Kybernetes, Vol. 3 Iss: 1, pp.17 - 25ე

[4]. M.F.Porter, An algorithm for suffix stripping. Program, v.14, no. 3, pp 130-137, July 1980

[5]. Harris A. C., Georgian Syntax: A Study in Relational Grammar. Cambridge University Press, Apr 30, 2009 - Language Arts & Disciplines - 352 pages

[6]. M.Khachidze, M.Tsintsadze, M.Archuadze, G.Besiashvili concept pattern formation in semantic search problems- GESJ: Georgian Electronic Scientific Journals, Computer Sciences and Telecommunications. Pp:13-20 2014

[7]. M.Khachidze, M.Tsintsadze, M.Archuadze, G.Besiashvili. Complex System State Generalized Presentation Based on Concepts. In: Application of Information and Communication Technologies (AICT), IEEE 8th International Conference . Kazakhstan, Astana 15-17 Oct. 2014 pp:1-4

[8]. N. Chinchor, MUC-4 Evaluation Metrics, in Proc.of the Fourth Message Understanding Conference, pp. 22–29, 1992.

[9]. Salton G., Buckley C. (1988), “Term-weighting approaches in automatic text re trieval”, Information Processing and Management, vol. 24 (5),pp. 513–523

[10]. Davis R. And Lenat D. (1982), Knowledge-Based Systems in Artificial Intelligence. McGraw-Hill Advanced Computer Science Series.

[11]. Egozi O., Markovitch S. and GabrilovichE. (2011), “Concept-Based Information Retrieval using Explicit Semantic Analysis”, ACM Transactions on Information Systems, Vol. 29, No. 2, Article 8, Publication date: April 2011.

[12]. Gabrilovich, E. and Markovitch, S. (2007), “Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In 20th International Joint Conference on Artificial Intelligence (IJCAI’07) procedings of international conference in Hyderabad, India, January 6-12, 2007, Morgan Kaufmann Publishers, pp. 1606–1611.