Published in the USA
European Journal of Technology and Design
Issued since 2013.
E-ISSN: 2310-3450
2023. 11(1): 9-14

DOI: 10.13187/ejtd.2023.1.9
https://ejtd.cherkasgu.press

Ontologies in Information Retrieval

Nikita S. Kurdukov [a, *]

[a] Department of Informative Studies, Institute of Information Technologies, Russian Technological University (RTU MIREA), Moscow, Russian Federation

Abstract

The article explores information retrieval technologies. The difference between information retrieval and information retrieval is shown. The search for information includes: information retrieval, semantic retrieval, and ontological retrieval. The problems of existing information retrieval technologies are described. Nine reasons for the inadequacy of information retrieval are described. A brief systematics of information retrieval methods is given. Current trends in the development of information retrieval are described. The article proves that the existing technologies of information retrieval are morphological. Work in the field of semantic search has led to the search for semantic information, but has not led to the creation of semantic search technology. The concept of complete information retrieval, which includes the search for morphology, content and ontology, has been introduced. The problems of the development of semantic search are described. The paradigms of informational, semantic and ontological search are given. It is shown that information retrieval is one-level and morphological. Semantic search is two-level. Ontological search is multi-layered. The key parameters of sematic and ontological search are highlighted: terminological relations, meaning of meaning, concept, knowledge. A search alternative is marked: the alternative is either a short time and a high volume of results, or a long time and a smaller volume of search results.

Keywords: information set, morphological search, semantic search, ontological search, content.

1. Introduction

The number of data in the Internet is growing exponentially (Azad, Deepak, 2019). Then the reflection. For example, problems in Big Data (Levin, Tsvetkov, 2017; Hariri et al., 2019). Limited growth of information is outpacing the growth in the number of methods to extract desired information (Azad, Deepak, 2019). Informational search is now the main tool for extracting information (Guo et al., 2020) in the network and in information storage systems. Informational network search doesn't yield adequate results in a row reasons. The first cause is polysemy. It consists in the fact that search patterns, or keywords submitted by a user can relate to multiple topics. In result polysemy search results can be are not focused on the topic of interest.

The second reason for the inadequacy of search results is the presence of information uncertainty. Information uncertainty is the standard state of search. In scientific research, search information is always known approximately. The research is characterized by information uncertainty. Information uncertainty at the beginning of a search leads to inaccuracy or irrelevance of search results.

* Corresponding author
E-mail addresses: nskurdyukov@gmail.com (N.S. Kurdukov)

The third reason for search inaccuracy is the length of the query and has to do with time. The query may be too short for the search engine to properly understand the meaning of what the user is looking for. The reason is subjective. The shorter is the query, the faster is the search. In practice, the average size of a web search is 2-4 words. The fourth reason for the inaccuracy of information retrieval is cognitive. It is caused by the lack of confidence and competence of the user. Under these conditions, the user is often unsure of what he is looking for until he sees the results. Even if the user knows what they are looking for, they don't always know how to compose correctly Request.

The fifth reason is the poor use of semantic relations and the poor use of auxiliary terms. Misuse of relationships distorts the meaning and renders the results of an informational slip useless.

The sixth reason for the inadequacy of the search is that information retrieval, which is morphological, is widely used. As a result of the search, an information set is formed according to morphological, not semantic features.

The seventh reason for the inadequacy of information retrieval is that the search methods do not take into account and do not assess the factor of information uncertainty.

The eighth reason for the inadequacy of information retrieval is that the overwhelming majority of search methods, with the exception of artificial neural network methods, do not use the ontological approach and the method of ontologies.

The ninth reason for the low efficiency of information retrieval technologies is the lack of methods that use different methods that take into account the criteria of "correspondence to meaning".

These reasons motivate the improvement of existing search methods, the development of new methods and new models of information retrieval. Such new methods include models of ontological search.

2. Discussion and results

Features of information retrieval.

Information retrieval as a technology is referred to the field of applied informatics (Polyakov, Tsvetkov, 2002) and is classified as a specialized information technology. Specialized information retrieval systems are used for information retrieval (IR). Retrieving information is a must for many applications, such as scientific research, dissertation research, digital library work, expert search, web search, etc.

Analysis of publications in the field of information retrieval indicates a growing trend of diversification of information retrieval methods. A significant part of the work is not integral. This is due to the fact that in many literature sources there are no clear requirements for identifying evidence of the truth of the information found. In information retrieval, the methods of correlation analysis (Tsvetkov, 2012), oppositional analysis (Tsvetkov, 2014a), dichotomous analysis (Kudzh, 2017) are not used. Therefore, the generalization of methods and the theory of information Searches are currently a challenge.

Existing models of neural information retrieval were often studied under homogeneous and narrow conditions, which significantly limited the understanding of their application for heterogeneous information (Thakur et al, 2021). Most web-based information search queries fall into the following categories (Azad, Deepak, 2019).

1. Information queries covering a broad topic, for which there may be thousands of alternative results.

2. Information requests covering a narrow topic that cannot be disclosed within the scope of the request, but can be disclosed by auxiliary iterative requests.

3. Navigation queries: Queries that search for a specific website or URL.

4. Transactional requests: Requests that demonstrate the user's intent to perform a specific action.

The first and second points are dominant. They are characterized by information uncertainty and the absence of semantic search criteria.

At present, the results of information retrieval are processed mainly using indexes and ontologies. At the same time, the use of ontologies in queries is not practiced. The use of ontologies is based on exact matches and is hidden from users. The use of morphological queries leads to the problem of terminological ambiguity (Tikhonov i dr., 2009). Morphological queries and search index are not based on the same set of terms. This is also known as dictionary problems (Furnas et al, 1987). Deficiencies in information retrieval technologies motivate the transition to new

methods. One of the new methods is the advanced query method (Azad, Deepak, 2019). It also uses conditional relevance feedback. This idea is to incorporate user feedback into the search process to improve the end result. In particular, the user provides feedback on the received documents in response to the initial request, indicating the relevance of the results. This idea is based on the inclusion of the cognitive space of a person in the space of information search. The main task of the search is the correspondence of meaning, but it is not yet explicitly designated. If we detail this task, we get a scheme of complete information retrieval.

Morphology-semantics-ontology.

The classical search is morphological, that is, it is built by searching for matches of morphological forms. It can be considered as the first and not the final stage of the search.

Search for semantic information

The search for semantic information is the second stage of complete information retrieval. There are works on semantic search (Raphael, 1964; Pejtersen, 1998). This was the work of the early years, when researchers naively thought they were doing a semantic search. But analysis shows that this is what was called "semantic search", but the search for semantic information (Amati, van Rijsbergen, 1998). The search for semantic information as an object and "semantic search" as a technology belong to different categories and cannot be identified. Therefore, the original direction of semantic search is now more accurately defined as the search for semantic information (Chebil, Soualmia, 2023).

A number of papers (Amati, van Rijsbergen, 1998) have attempted to use "semantic information theory" (SIT). This theory was created as an alternative to the information theory laid down by the works of C.E. Shannon. The SIT has not been finalized. Its interpretation is not definite. Vaguely: It was built very broadly and vaguely. SIT relied on research in the field of polysemantic logic and philosophy, but not on formalism in the field of computer science. And the use of the term "information" was used to denote the description and transfer of this description from one subject/object to another (Amati, van Rijsbergen, 1998). General the difference between SIT and C.E. Shannon's theory of information is that information is conveyed not by an ordered sequence of binary symbols, but by means of a formal or natural language in which logical statements are defined and explained by semantics. However, these ideas have not been implemented to this day and have remained as wishes. They are useful for semantic search. However, it should be noted that in reality there is a search for semantic information.

The ontological approach to information retrieval should be noted as a promising direction (Mustafa et al, 2008). Semantic methods of information retrieval must understand the meanings of the concepts that users specify in their queries. However, the main drawback of existing methods of semantic information retrieval is that none of them takes into account the context of the concept (Mustafa et al, 2008). To solve this problem, the approach of thematic similarity is used. It is used to search for information to capture the context of a particular concept. In addition, source metadata in the form of RDF triples is used.

The concept of relevance is a hot topic in the process of searching for information. In recent years, the dramatic growth in the number of digital documents has highlighted the need for new approaches and more effective methods to improve the accuracy of IR systems to meet the information needs of real users to measure the semantic relationship between words. This approach is based on ontologies presented using a common knowledge base to dynamically build a semantic network. This network is based on linguistic properties and, when combined with a metric, creates a measure of semantic connectivity.

The problem of semantic search in biomedical digital libraries is described in (Ebeid, Pierce, 2021). It uses a vector approach to search. It describes a method based on knowledge graph embedding, which provides semantic relevance search and ranking of biomedical literature indexed in PubMed.

Chebil and Soualmia (2023) provide a relatively complete approach that includes a query extension technique. The approach proposed in this study combines probabilistic networks (PN), vector space model (VSM), and pseudo relevance feedback (PRF) to evaluate and add relevant concepts to the user's original query index. First, query extension is done using PN, VSM, and domain knowledge. Then, in the second step, PRF is used to enrich the query user using the same approach used in the first phase of the extension. To evaluate the performance of the developed system, called the Conceptual Information Retrieval Model (CIRM), several query extension

experiments are conducted. Experiments have shown that the use of two measures of possibility and necessity in combination with cosine similarity and PRF improves the process of information retrieval. In all of these methods, such a search is not complete. It excludes the issues of finding latent information and tacit knowledge (Bolbakov, 2016).

Basics of semantic search.

Semantic search is based on the idea of semantic environment (Tsvetkov, 2014b) and semantic modeling. The ideas of semantic modeling go back to the work of Carnap (Carnap et al., 1953) and Luciano Floridi (Floridi, 2004). Carnap's works can be interpreted as wishes: "what should be and what would be desired". A more fundamental approach is proposed by L. Floridi. He introduces the concepts of "Strict Semantic Information Theory" (TSSI) and "Weak Semantic Information Theory" (TWSI). TSSI is based on truth values, not probability distributions. TWSI is based on probability distributions and actually describes C.E. Shannon's theory of information. There is a paradox, also revealed in (Tsvetkov, 2014c), between content and information volume. Floridi (2011) examines the relationship between "Semantic Information and Theory and Correctness of Truth". Floridi associates probabilistic characteristics with semantics, which is conditional and limits the theory. The main inaccuracy of L. Floridi in the Interpretation of the Concept of Truth. In the actual practice of "Truths a" there is a conditional concept. For a long time, the world was described by a geocentric system. This was believed to be true. But scientific research has led to a different model of the world, the heliocentric one. These models and the truths based on them contradict each other. Another example is the geometry of Euclid and Riemann. These models are not consistent, but complement each other. Semantic information is there But epistemologically, semantic information does not change and does not depend on interpretation or truth criteria

With regard to semantic information theory or complete information theory, one can agree with the opinion (Zhong, 2017). "Information (an information model) that is truly useful to people should consist of three components: a form called syntactic information, a meaning called semantic information, and a utility called pragmatic information" (Zhong, 2017). A fourth component, the ontological, should be added here. The "Information" term is amorphous. A more accurate term in the field of information retrieval is "information model". An information model has integrity, limitation, and structure. The ontological component of the information model is that it must conform to generally accepted concepts and contain particular and general knowledge. Concepts and general knowledge are ontological factors.

The quote above allows us to move on to the morphology of information models (Jeulin, 2021). The topic of morphology is still considered separately from the theory of information and from the theory of information modeling.

Paradigms of information and ontological search.

Information search is the simplest, but it is also divided into categories. It is based on structural information units (SIU), patterns (P), information set (IS), information clusters (IC), comparison methods (CmM), and search results set (SSR).

$$SIU \rightarrow P \rightarrow IS \rightarrow IC \rightarrow SSR \ (1)$$

Paradigm (1) is interpreted as "one pattern – one set of searches". Paradigm (1) has two implementations for complex patterns P(A, B)

$$SIU \rightarrow P(A, B) \rightarrow IS \rightarrow IC \rightarrow (SSR(A) \cup SSR(B)) \ (2)$$

Paradigm (2) is interpreted as follows: "one pattern – several sets of search results". Paradigm (2) is found in simple search engines, such as searching for files in the Windows operating system. If the search pattern includes two words, then all the words in the pattern are searched independently. The search result consists of a collection of sets for each word in the pattern. This search method takes little time but creates large amounts of information that the user must analyze on their own. The load is transferred to the cognitive area of the person.

Another search paradigm takes into account the terminological relationships (R) between the words of the pattern. For example, consider two words as in paradigm (2)

$$SIU \rightarrow P(A, R, B) \rightarrow IS \rightarrow IC \rightarrow SSR(A, R^*, B) \ (3)$$

Paradigm (3) is interpreted as follows: "one pattern with relations – one set of searches with reduced relations (R*)". Paradigm (3) is found in the search engines of the word processor Word. If the search pattern includes words and relationships, the search result contains a simplification or

modification of the relationship, but with the words included. Paradigms (1-3) describe morphological search.

To implement semantic search (SR), you need to specify a meaning value (MM). The principal difference between semantic search is the presence of two levels of search.

$$SIU \rightarrow P(A, R, MM, B) \rightarrow IS \rightarrow IC \rightarrow SSR(A, R^*, B) \quad (4)$$

$$SSR(A, R^*, B) \rightarrow SmS\ (A, R^*, MM^*, B) \quad (5)$$

The first level of search (4) is morphological. The second level of search is semantic. The result of the search is a set of meanings that do not exist in paradigms (1-3). The result of semantic search (5) is a semantic set (SmS). A feature of expression (5) is that the value of the meaning at the start of the search (MM) may differ from the value of the meaning at the search result (MM*). The criterion for the relevance of semantic search is the relationship

$$MM \approx MM^* \quad (6)$$

Ontological search (OR) differs in the number of levels and the result of the search.

$$SIU \rightarrow P(A, R, MM, B) \rightarrow IS \rightarrow IC \rightarrow SSR(A, R^*, B) \quad (7)$$

$$SSR(A, R^*, B) \rightarrow SmS(\ A, R^*, MM, B) \quad (8)$$

$$SmS(\ A, R^*, MM, B) \rightarrow SO\ (C, Kn, R^{**}, MM^*) \quad (9)$$

Ontological search contains the first level of morphological search (7), the second level of semantic search (8), and the third level of ontological search. The results of morphological typing and semantic search are commensurate because they describe the same objects with different completeness. The results of the ontological slip (9) and the morphological search are qualitatively different, since in the ontological search we find not objects, but: concepts (C), knowledge (Kn), generalized relations (R**), generalized meaning (MM*). All generalizations go beyond a single object and describe a group of objects. Ontological search is based on semantic search, correspondence of meaning, and conceptual modeling. Expressions (8) and (9) can have sublevels. Therefore, the scheme (7-9) is multi-level.

3. Conclusion

The problem with all types of searches is the complexity, search time, and volume of search results. Complexity reduces search time. In search engines, there is an alternative, either a short time and a high volume of results, or a smaller volume but a longer search time. Ontological search has a greater number of levels of search and analysis. The semantic and ontological levels include analysis as part of search. The conducted research gives grounds to introduce the concepts of "morphological search", "semantic search", "ontological search". There is reason to consider the existing information search to be morphological. Morphological factors play a major role in it. Semantic search involves semantic analysis. Ontological search involves generalization and conceptual analysis. All types of search reduce information uncertainty. Morphological search is the simplest because it uses a well-formalized space of parameters. For informational, semantic, and ontological models, morphology determines their representation. With semantic search factors are little used in search technologies as well. Orthology can define the structure of a model or object. The semantics of information models is determined by their content and relation to reality. The relation to reality determines the conditional truth. Summing up, it should be stated that the concepts of information retrieval and information retrieval are not identical. Information retrieval is one technology with one level of retrieval. Information retrieval includes different technologies with a large number of levels of search and analysis. Searching for information yields results that fall into different categories.

References

Amati, van Rijsbergen, 1998 – *Amati, G., van Rijsbergen, K.* (1998). Semantic information retrieval. Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information. Boston, MA: Springer US. Pp. 189-219.

Azad, Deepak, 2019 – *Azad, H.K., Deepak, A.* (2019). Query expansion techniques for information retrieval: a survey. *Information Processing & Management*. 56(5): 1698-1735.

Bolbakov, 2016 – *Bolbakov, R.G.* (2016). Tacit Knowledge as a Cognitive Phenomenon. *European Journal of Technology and Design*. 1(11): 4-12.

Carnap et al., 1953 – *Carnap, R., Bar-Hillel, Y., Cambridge, M. I. T., Dretske, F.* (1953). Semantic theories of information. *Philosophy of Science*. 4(14): 147-157.

Chebil, Soualmia, 2023 – *Chebil, W., Soualmia, L.F.* (2023). Improving semantic information retrieval by combining possibilistic networks, vector space model and pseudo-relevance feedback. *Journal of Information Science*. P. 01655515231167293.

Ebeid, Pierce, 2021 – *Ebeid, I.A., Pierce, E.* (2021). MedGraph: An experimental semantic information retrieval method using knowledge graph embedding for the biomedical citations indexed in PubMed. *arXiv preprint arXiv*. 2112.06348.

Floridi, 2004 – *Floridi, L.* (204). Outline of a theory of strongly semantic information. *Minds and machines*. 14: 197-221.

Floridi, 2011 – *Floridi, L.* (2011). Semantic information and the correctness theory of truth. *Erkenntnis*. 74: 147-175.

Furnas et al, 1987 – *Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.* (1987). The vocabulary problem in human-system communication. *Communications of the ACM*. 30(11): 964-971.

Guo et al., 2020 – *Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., ... , Cheng, X.* (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management*. 57(6): 102067.

Hariri et al., 2019 – *Hariri, R.H., Fredericks, E.M., Bowers, K.M.* (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*. 6(1): 1-16.

Jeulin, 2021 – *Jeulin, D.* (2021). Morphological models of random structures. Cham: Springer.

Kudzh, 2017 – *Kudzh, S.A.* (2017). Dihotomicheskij strukturnyj analiz [Dichotomous structural analysis]. *Slavyanskij forum*. 2(16): 7-11. [in Russian]

Lyovin, Tsvetkov, 2017 – *Lyovin, B.A., Tsvetkov, V.Ya.* (2017). Informacionnye processy v prostranstve «bol'shih dannyh» [Information processes in the space of "big data"]. *Mir transporta*. 15. 6(73): 20-30. [in Russian]

Mustafa et al, 2008 – *Mustafa, J., Khan, S., Latif, K.* (2008). Ontology based semantic information retrieval. *2008 4th International IEEE Conference Intelligent Systems*. IEEE. T. 3. P. 22-14-22-19.

Pejtersen, 1998 – *Pejtersen, A.M.* (1998). Semantic information retrieval. *Communications of the ACM*. 41(4): 90-92.

Polyakov, Tsvetkov, 2002 – *Polyakov, A.A., Tsvetkov, V.YA.* (2002). Prikladnaya informatika [Applied informatics]. Moskva: Yanus-K, 392 p. [in Russian]

Raphael, 1964 – *Raphael, B.* (1964). SIR: A computer program for semantic information retrieval. MAC-TR2 Project MAC MIT June.

Rinaldi 2009 – *Rinaldi, A.M.* (2009). An ontology-driven approach for semantic information retrieval on the web. *ACM Transactions on Internet Technology (TOIT)*. 9(3): 1-24.

Thakur et al., 2021 – *Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.* (2021). Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv*. 2104.08663.

Tihonov i dr., 2009 – *Tihonov, A.N., Ivannikov, A.D., Tsvetkov, V.Ya.* (2009). Terminologicheskie otnosheniya [Terminological relations]. *Fundamental'nye issledovaniya*. 5: 146-148. [in Russian]

Tsvetkov, 2012 – *Tsvetkov, V.Ya.* (2012). Framework of Correlative Analysis. *European researcher*. 6-1 (23): 839-844.

Tsvetkov, 2014a – *Tsvetkov, V.Ya.* (2014). Opposition information analysis. *European Journal of Technology and Design*. 4(6): 189-196.

Tsvetkov, 2014b – *Tsvetkov, V.Ya.* (2014). The Semantic environment of information units. *European researcher*. 6-1 (76): 1059-1065.

Tsvetkov, 2014c – *Tsvetkov, V.Ya.* (2014). The K.E. Shannon and L.Floridi's amount of information. *Life Science Journal*. 11(11): 667-671.

Zhong, 2017 – *Zhong, Y.* (2017). A theory of semantic information. *China communications*. 14(1): 1-17.