Semantic Web-Based Information Retrieval Models: A Systematic Survey

Anil Sharma[1(✉)] and Suresh Kumar[2]

[1] USIC&T, Guru Gobind Singh Indraprastha University, Delhi, India
anilsharma.iimt@gmail.com
[2] Ambedkar Institute of Advanced Communication Technologies & Research, Delhi, India
drsureshpoonia@gmail.com

Abstract. Effective representation of semantics in information sources has been central to Semantic Web (SW), since its inception. An information retrieval (IR) system must exploit the semantic knowledge embodied in web resources. Several attempts were made by researchers to make retrieval systems capable of utilizing web semantics. As a result, IR systems exploiting Semantic Web technologies were proposed in literature. In this paper, we have presented various intelligent models for information retrieval on SW. Our work mainly focuses on systems based on multi-agent, ontology, soft computing and concept-based paradigm employed for information retrieval on SW. Some existing surveys tried to comprehend various intelligent information retrieval models on SW, but their scope is limited. In this paper, we are providing a systematic and comprehensive elucidation of various intelligent information retrieval models with their basic approaches, key features and limitations in the context of SW. We have also provided a comparison of reviewed IR models for critical analysis.

Keywords: Information retrieval models · Semantic Web · Agent-based model · Concept-based model · Ontology-based model · Soft Computing-based Model

1 Introduction

With the proliferation of web technologies, huge amount of information is uploaded to and downloaded from the Internet every day. Due to information overload problem, users are struggling to get relevant information on the web. Search engines perform their work based on keyword search, making it difficult to tackle the problem of synonyms and polysemy. This makes the existing information retrieval methods ineffective with low precision and recall rates. This problem occurred as world wide web was not intended to be processed by machines. Although, the web page includes metadata and actions to be taken, but it does not provide interpretation of semantic of contents.

Tim Berner Lee proposed the solution of this problem in terms of the Semantic Web (SW) [1]. The word semantic simply means meaning. Meaning of data provides effective usage of data by establishing relationship and context with other data items.

U. Batra et al. (Eds.): REDSET 2019, CCIS 1230, pp. 204–222, 2020.
https://doi.org/10.1007/978-981-15-5830-6_18

SW provides semantics and context to web resources, enabling machines to understand the meaning of web contents. With the advent of SW, machines can process web contents more intelligently that aid in effective information search and retrieval. Ontology is a central concept in SW. Resource Description Framework (RDF), Web Ontology Language (OWL) and SPARQL are fundamental technologies behind the success of SW.

Ontology is taxonomy of domain concepts represented in the form of entities, their attributes and the relationships between entities. It is regarded as knowledge representation tool that formulate concepts of a domain in SW [2]. For representation of ontology concepts embodied in web resources, RDF was proposed in SW framework [3–5]. Later, to generalize the process of representation, processing and inference of web contents, OWL was developed under SW framework. OWL resulted as a stronger language with better machine interoperability as compared to RDF. DAML + OIL are universal SW markup language providing machine with capability to read, interpret and infer data. In [6], DAML + OIL was used for precise knowledge representation and retrieval.

WordNet [7] represents a general purpose ontology that stores words with their synonyms. WordNet API is employed in query expansion in [8, 9]. Wikipedia is a web-based domain independent structured encyclopedia [10]. Wikipedia found to be fruitful in applications such as question answering, named entity disambiguation, text categorization and computing document similarity. Wikipedia is also employed for computing semantic similarity between query and web document concepts [11]. Jena is a semantic framework used to manipulate RDF data. RDF data is manipulated using Jena tool [3, 9]. This tool is a Java API that employs SPARQL query language for processing, retrieval and manipulation of data in RDF format [9, 12].

Development of web technologies and universal availability of web based information systems are main reasons behind the existence of web based information retrieval (IR) systems. Despite of years of research in the IR field, still there is no proposal that wins. Effective utilization of web semantics by machines was one of the ideas behind conceptualization of SW. As a result, IR systems exploiting SW technologies were proposed.

1.1 Motivation of the Survey

It has observed from literature that since last two decades researchers have been very interested in SW based information retrieval. But despite this fact only a few surveys [13–19] were published on the topic. Although these surveys contributed well in the field but were limited in scope due to lack of comprehensive approach in coverage of topic.

A survey carried out in [13] focused only on agent-based personalized semantic IR. This survey provided description of only personalized semantic search based on mining techniques, neural networks, genetic algorithm, ontology and collaborative filtering, but failed to include comparative analysis of frameworks cited in the survey. The survey in [14] concentrated on concept-based IR models on SW. Some frameworks on

WordNet, SW, conceptual indexing and Word Sense Disambiguation were also discussed in this survey, but very limited frameworks were included in this survey. The survey in [15] discussed soft computing based intelligent IR models. Further, application of probability theory, Fuzzy Logic, Genetic Algorithms and Artificial Neural Networks in context of soft web mining were also deliberated. Inclusion of few numbers of frameworks in this survey made it limited in scope. The survey in [16] explored SW and IR architecture along with some prototype systems for query expansion. Again, discussion included few frameworks mentioned without challenges and research gaps. The survey in [17] enlisted various ontology based and agent-based models for SW search along with brief description about techniques used in cited frameworks. This survey included few related frameworks that limit the scope of this survey. The survey in [18] included a brief description of agent based and ontology-based IR models on SW but failed to include research gaps for enlisted frameworks. Moreover, the survey in [19] employed classification parameters for comparison of semantic search engines approaches. A comparison of cited approaches was also presented in this survey. But only few proposals were discussed without exhaustive coverage of state-of-the-art frameworks.

It is evident that referred surveys lack comprehensive coverage of SW-based IR models. In order to understand different intelligent IR models on SW one need to refer many scattered sources. Motivated with this fact, we decided to present a systematic and comprehensive survey on intelligent IR models on SW. A comparative analysis of referred surveys with our survey based on attributes (taxonomy, comparative analysis, tabular representation, graphical representation and research directions) is presented in Table 1, where (✓) shows inclusion and (×) denotes non-inclusion of above said attributes.

Table 1. Comparative analysis of referred surveys with present survey.

Survey Papers	Models Taxonomy	Comparative Analysis of Frameworks	Tabular Representation	Graphical Representation	Research Directions
Thangaraj et al. [13]	×	×	×	×	✓
Ali and Ahmed [14]	×	×	×	×	×
Ahmed and Ansari [15]	×	×	×	×	×
Singh and Jain [16]	×	×	×	×	×
Sharma A. [17]	×	✓	✓	×	×
Balan et al. [18]	×	✓	✓	×	×
Ezhilarasi et al. [19]	×	✓	✓	×	✓
Present survey	✓	✓	✓	✓	✓

1.2 Scope and Organization of the Survey

The scope of our survey includes:

- Explanation of intelligent IR models on SW environment.
- Tabular presentation of IR models on SW for deeper understanding.
- A comparative analysis of present study with cited studies focusing on basic approach, methodology, techniques and limitations is also presented.

The present survey is organized as follows: Introduction is included in Sect. 1, while Sect. 2 discusses web-based information retrieval. Section 3 throws light on traditional information retrieval models. Section 4 discusses intelligent information retrieval models, in which we focused mainly on concept-based model, agent-based model, ontology-based model and soft computing-based model for Intelligent IR. Section 5 is about discussions and analysis. Conclusions and research directions are presented in Sect. 6.

2 Web-Based Information Retrieval

An IR system is responsible for pre-processing; organizing, storing and indexing information for retrieval of relevant documents in response to user's query, while query-document matching algorithm and relevance ranking of resultant documents being most critical activities during retrieval process. A general model of information retrieval system is discussed in [20]. With the revolutionized development of online Information Systems in every domain, web-based IR has become important area of research. Dynamic nature of Web makes web-based IR systems different from traditional IR systems in terms of knowledge representation, indexing, query expansion and interpretation, retrieving relevant documents, ranking and presentation of resultant web pages.

Information retrieval models are broadly divided into two categories i.e. Traditional IR and Intelligent IR. First, traditional IR models were based on keyword search and were mainly dependent upon syntactics of search terms. These systems suffered mainly due to two reasons: first problem of synonyms and polysemy; and second, lack of standards for information representation. Semantics of search terms were ignored in traditional search methods as they focused on syntactic properties of search words. Second in Intelligent IR, with the conceptualization of semantic web, inclusion of semantics of keywords was realized. Now, more metadata can be embodied in knowledge base regarding a keyword. Information becomes machine understandable rather than just machine readable which paved road for intelligent models in information retrieval. Literature shows that in the last two decades researchers have proposed information retrieval systems based on various models which dealt with problems of traditional IR models. In Fig. 1, hierarchal representation of IR models is shown.

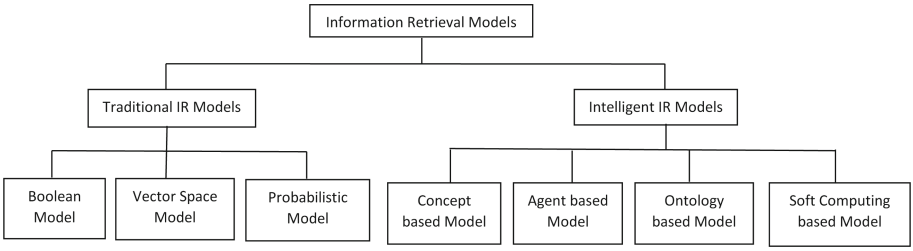

Fig. 1. Hierarchal representation of Information Retrieval Models

3 Traditional Information Retrieval Models

Traditional IR models were based on keyword search and were mainly dependent upon syntactics of search terms. Three models based on traditional IR are explained below as:

3.1 Boolean Model

In Boolean model, web pages and user's queries are denoted as index terms and Boolean expressions on index terms. This model exploits classical set theory and Boolean expressions to denote web pages and users query [21].

Limitations: Boolean Model has two major shortcomings: first, it is not able to deal with partial matching of documents with search terms as this model provides results based on exact match between web pages and user's query. Moreover, the fetched web page ranking is not considered in this scheme. Second, it's not easy to model each search query into Boolean expression.

3.2 Vector Space Model

Vector Space Model (VSM) is based on vector representation of documents and queries in multi-dimensional space. Non-binary weights based on term frequency are assigned to key terms in search query and web documents and the degree of similarity is calculated based on these weights [22].

Limitations: The problem with VSM is that it is unable to establish relationships between key terms making them unable to link with each other. Further, the relevance ranking of fetched web pages was not considered in this scheme.

3.3 Probabilistic Model

This model uses probability theory as underlying principle. Let Q represents user's query and W represents set of web pages while $S \subseteq W$ is a subset of web pages that consists of relevance information related to web pages and user's query [23, 24]. In this model, relevance ranking of fetched web pages depends upon descending order of probability of belongingness to subset *S*.

Limitations: Drawback of this model is that it does not incorporate frequency of key terms in relevance calculation. Also, some factors affecting relevance judgment like user preferences and score based on web page relevance to other users were not considered in this approach.

4 Intelligent Information Retrieval Models

There have been significant researches in devising intelligent methods for information retrieval on SW. Methods using various techniques like concept-based, agent-based, ontology-based and soft computing techniques etc. have been reported in literature for making machine more intelligent and web more machines readable. Following are some intelligent models used for information retrieval on SW:

4.1 Concept-Based Information Retrieval Model

Users recognize information in terms of concepts but mostly information retrieval systems employ keyword-based search mechanism. Concept can be considered as a collection of terms that together identify the clear meaning of the intended context [25]. The Concept-based IR system provides results based on conceptual relationship between web pages to the terms in user's query, rather than based on literal meaning or context found in web pages. IR system should be intelligent enough to capture the search intent and conceptual meaning of search terms in query. Conceptual Indexing and Word Sense Disambiguation (WSD) are two such approaches. Conceptual Indexing system automatically extracts conceptual information from documents and build hierarchical concept graph dynamically [26]. WSD hunt for sense implied by a term and based on that sense it assigns context to the term, whereas sense is defined by concepts. Concept-based system presents additional intelligence to IR by using ontologies.

As a result, Concept-based IR systems can fetch relevant documents even if query terms and their synonyms are not present in documents because the retrieval is guided by semantics not just by syntactic properties of terms in search query. A concept-based IR model is discussed in [27]. In this model, both information resources and user query both are represented as concepts. Word Sense Disambiguation (WSD) is employed to tackle problem of one word representing many concepts and to identify its context. Concept similarity is performed to find the conceptual match between user query and documents in information resources on Web. Knowledge repository plays an important role to facilitate concepts and their relationships with other concepts.

Limitations: The main drawback is that limited terms are incorporated in ontologies and conceptual information augmented about these terms is not complete.

In [28] authors proposed a concept-based information retrieval model that uses corpus such as Wikipedia for finding term co-occurrences and relationships between concepts. This approach employs Explicit Semantic Analysis (ESA) [29] which treats text semantics (meaning) as a combination of concepts found in knowledge resource rather than just depending upon syntactic structure of text. Feature selection works on query concepts to make it optimized for reasoning and matching. This method uses

Wikipedia as a knowledge resource to extract concepts embodied in it. So, this method cannot be deployed in very domain as it requires a pre-existing knowledge resource.

In [30] an improved concept-based IR system was presented. This model uses term clustering technique to find proximity of search term in a document. Weights and penalties were assigned to nouns using Okapi weighting scheme. Semantic frame provides matching between query and documents. Model may need significant modifications before deploying into more specialized domains like medical as concepts are overlapping in multiple clusters.

In [31] an IR system based on fuzzy formal concept analysis (FFCA), concept hierarchies and automatically built domain ontologies is proposed. Concept lattice is generated by exploiting context information from syntactic relations of a term with most frequent verbs in corpora. Fuzzy formal concept is used for relevance match between user query and search documents. Other fuzzy relations (like resemblance and tolerance) may be employed in fuzzy relational ontological modeling.

In [32] authors proposed concept-based semantic search in cloud employing Wikipedia ontology, double score weighting formula and Semantic searchable encryption scheme. Weighting scheme considered higher preference for concept associated with general meaning over concept associated with higher frequency terms of a document. Performance issue needs further investigations as encrypted search is not very efficient.

4.2 Agent-Based Information Retrieval Model

Agents are computer programs, belong to the field of Artificial Intelligence (AI) that learn the pattern and behavior of a user and act on his behalf. These agents implement the web retrieval service based on ontology [4]. Agents consider user's background, web knowledge, user's interests and searching patterns to satisfy their information need automatically. In literature, various agent-based models for information retrieval are reported. Each model used different types of agent depends specific objective. By employing ontologies, IR systems gain retrieval performance, but personalization and degree of relevance show no significant improvement [33].

A multi agent-based IR model is presented in [34]. This agent-based IR system employs multiple agents for a specific task in retrieval process. User agent consists of inference engine along with learning mechanism and environment module. Information gathering agent comprises of search strategies and optimization module. Semantic extraction module is used to extract semantics (nature, structure and relationships) of user's query and web pages. Semantic matching between user's query and web pages is performed by semantic matching agent and relevant results are forwarded to user agent.

Limitations: Building, managing and updating knowledge repositories are challenging tasks that require additional efforts.

In [4], authors proposed a multi-agent based intelligent information retrieval model in semantic web. In proposed work application of Information Collection, Storing, Reasoning, and Query Agents are reported for different task. Moreover, the proposed framework uses Resource Description Framework (RDF) model for web resources description, modeling, and web resources content representation. In the same paper, Web Ontology Language (OWL) was used to construct domain ontologies, which

provide knowledge base for reasoning semantic query. This model makes use of four agents: Information Collection Agent is employed for metadata extraction from web page content description in semantic web. Storing Agent is used for storing metadata coming from information collection agent. Reasoning Agent perform semantic reasoning (semantic matching and keyword retrieval from semantic relevance). Query Agent provides the results to meet user's requirement by querying the metadata based on semantic ontology. Hence the application of multi agent improved the efficiency and precision.

In [33], authors applied intelligent agents to retrieve the information in semantic web. Further, authors exploit ontology knowledge to speedup query processing and improve accuracy. The presented model uses following agents: User Interface Agent which interact with user and make use of ontology knowledge to group user's information need. Retrieval Agent matches the user's requirement with resource description in database and collaborates with Management Agent to augment user's interest factor for arranging results in relevance of user's interest. Resource Description Agent captures the semantic description based on domain ontology from web pages captured by crawler and stores it into database. Ontology Collection Agent captures new ontologies and update existing ontologies by interaction of ontology base with www. Management Agent interacts with other agents for collaborating information of user's interest with Retrieval Agent for arranging results in user preference order. Matching module uses semantic matching algorithm for user's information request with semantic description in database.

In [35] authors proposed agent-based method for discovering web services. Semantic information related to web services plays an important role in their discovery by users. Using domain ontology and web ontology language (OWL) we can enhance the quality of representation of semantic information. Web service composition can be incorporated in the model for creating new web services by merging and reusing existing ontologies.

In [36] authors presented semantic web and agent based educational system that facilitates course contents and information in ontology form. Similarity between query concept and course resource is computed using least common super-concept (structural taxonomy based) similarity measure. A comparative analysis of Vector Space Model (VSM) and ontology-based IR indexing system is also presented in this paper. Feature-based methods for computing semantic similarity which provide additional knowledge about the concept and its relationship with other concepts were not considered in this proposal.

4.3 Ontology-Based Information Retrieval Model

Ontologies are knowledge representation tool and facilitate classification as well as mapping of concepts and their relationships in hierarchical structure. Literature shows employing conceptual knowledge (ontologies) in information retrieval process has contributed to solution of key limitations in information retrieval. An ontology-based IR system is presented in [37]. In this model, ontology vocabulary extraction is performed on query search terms. Vocabulary terms from documents are extracted based

on concepts of ontology and are presented as vector space. Similarity between query and document concepts is calculated using correlation matrix between concepts.

Limitations: Relationship between different concepts, inference of semantic information from concepts, unifying semantic representations and mapping of knowledge from heterogenous ontologies. Also, supervised learning approaches can be employed to identify semantic relations between query and document concepts.

In [12], authors presented an IR system based on domain ontology. To exploit semantic relationship between ontologies query language SPARQL is used in this proposal. Ontology provides concept hierarchy and logic reasoning support, which makes it suitable tool for semantic retrieval. In this model, SPARQL query language provides extraction of information by utilizing the association between concepts defined in ontology. As a future trend, the author gives an idea about using fuzzy ontology concepts for the proposed system. In [3], search engines based on Web Ontology Language was proposed. For indexing and retrieval of semantic relationships, an algorithm based on Web Ontology Language was proposed. Proposed indexing algorithm offers better ontology maintenance and retrieving algorithm facilitate better processing of user query. This scheme consists of these components: Repository of OWL-web pages, which was prepared by semantic web crawler automatically. Here author manually created this repository by preparing ontologies with Protégé ontology editor. The objective of Ontology Analyzer tool is semantic inference from web documents written in OWL. Authors applied Pellet Reasoner and Jena tool for this purpose. Thematic Repository is developed by Ontology Analyzer Tool, which acts as input for next component of proposed model i.e. Indexer, to store the index of ontology thematic repository. Retrieval scheme is employed for determination of relevance of information to end user using precision and recall. User Interface is used to obtain input query from user and to return relevant results.

In [9], authors presented an IR system based on domain ontology. Here meaning concepts are inferred from user's query. These inferred concepts and domain ontology are used for query expansion. SPARQL query is framed and used on knowledge base to return relevant web pages. The resulted pages are ranked according to query reference. Query expansion benefits the system by considering query concepts and synonyms of these concepts as well as new concepts associated to query. Query expansion exploits field ontology for finding terms related to original query. Semantic similarity between inferred concepts and domain ontology concepts is achieved using structure based measures [25].

In [38], authors conceived the idea of an intelligent IR system based on SW. In this model, semantic relations between web pages are estimated using proposed metric. Distributed Hash Table (DHT) is used for load balancing and range queries as well as for distribution and fault tolerance. In [6], authors introduced an IR system based on Web Ontology Language exploiting semantic markup. For documents and query semantic markup DAML + OIL SW language was used thus allowing inference at the time of document indexing, query processing and result evaluation. DAML + OIL allow reading, interpretation and inference to be done over the data by machines, making machines more intelligent in processing and retrieving information over the SW.

In [5], proposed SW based intelligent IR system for solving two problems. First is how to make web resources machine understandable? And second is how to implement domain knowledge concepts for semantic search? This system utilizes ontology for organizing metadata which not only provide contents of web resources but also provide semantic relationship documents and concepts hierarchy as a basis for semantic inference. Moreover, metadata and query are encoded in RDF. The results of user's query are sorted according to user's intent and presented in a suitable format.

In [8], an intelligent Cross-Lingual Information Retrieval (CLIR) system is presented that retrieved results also from web pages written in languages other than language in which user query is written. Results are returned in original query language after translation. Spell check technique is included to facilitate user with query support. It also supplements the search by including synonyms, related words of query tokens and semantic relations. This system uses query expansion and semantic relations to determine user's search intent and context to query thus reduces irrelevant web pages to be included in results set. Universal Networking Language (UNL) was employed for cross language support. The proposed search engine is tested for agriculture domain and performance is found reasonable with retrieval of relevant results.

In [39], authors proposed a knowledge retrieval process model using semantic metadata and artificial intelligence techniques. Based on ontology, metadata inside web document is queried which provide concept base retrieval in distributed environment of e-leaning resources. Proposed system employs ontology as vocabulary for Case-Based Reasoning (CBR) and ontology was developed using RDF language. Search returns both results based on semantic term and ontology concepts. The system is tested for information retrieval in digital library.

In [40], authors presented a model for personalized search engine using query clustering technique and SW, by utilizing context of query while considering search history. The keyword based search engines uses keyword matching approach yielding low quality results. Exploitation of popular search technique for page ranking method by some commercial entities to seek people attention has thrown a challenge to researchers in IR field. ECBR algorithm [26] is applied to estimate the degree of relatedness between query and service concept keeping synonym of query into consideration. This method was tested for service retrieval in transport domain ontology.

4.4 Soft Computing-Based Information Retrieval Model

IR systems suffer due to imprecise and vague knowledge representation in user query formulation. Soft computing is an effective tool to deal with such vagueness in information representation [41]. Literature shows that soft computing techniques such as fuzzy logic [42], rough set theory [43], artificial neural networks [44], genetic algorithms [45] and evolutionary computing [46] has been successfully applied in information retrieval. In [50] authors proposed a model of soft computing-based IR model in SW. Soft computing-based IR systems vary in semantic similarity measures used to match query and document concepts. These semantic similarity measures are broadly falling into three categories: taxonomical structure, feature, and information content based semantic similarity measures. Despite of advantages of each method none of these approaches clearly emerged as a solution.

Limitations: In taxonomical structure and Feature based methods similarity computation between two terms is based on ontological hierarchical structure and function of their properties (relationship to other similar terms in corpus), which is affected by degree of coverage of input ontology. Information content semantic similarity measure exploits an additional large text corpus to compute word frequency. So, this method cannot be deployed in every scenario, because existence of such a large corpus is not feasible for every domain. In our survey, we will restrict ourselves to fuzzy logic and rough set-based IR models.

In [47], an automated approach for annotating web services-based on fuzzy rules using VSM for semantic representation was introduced. Providing semantic annotation to web services helps them to get linked with relevant service concept in domain ontology. This facilitates automatic recognition, selection, retrieval, and composition of web services by machines. Fuzzy set theory is used to computer degree of membership function for calculating similarity between a service and service concept in domain ontology.

In [2], authors introduced three-layer architecture for traffic IR system, based on fuzzy ontology on SW. Authors used fuzzy linguistic variable ontology based concept relationships for semantic query expansion.

In [48] authors discussed the problem of incapability of domain ontology to deal with uncertain information due to lack of clear-cut boundaries between concepts of domains. The solution was provided in terms Fuzzy Ontology-based Intelligent IR system was proposed. The fuzzy ontology can deal with uncertainty of relations concept hierarchy of specific domain, thus providing more accurate results. The authors claimed that with their proposal the effectiveness of IR system improved significantly.

In [49] author implemented a concept similarity method featuring formal concept analysis and type-2 fuzzy sets. The proposal employs formal concept analysis (FCA) with many-valued context to address the problem of interval valued attributes of searched concepts. Concept similarity is proposed using FCA and fuzzy sets using Information Content (IC) approach. Limitation of this method is that knowledge contained in hierarchical structure of concept lattice (level and depth of concepts) is not included in performing semantic similarity measure.

In [51] authors proposed an ontology mapping framework based on rough set theory (RST) and concept lattice. Two ontological contexts were considered for construction of concept lattice. Similarities of two ontological nodes were measured using rough set approximations. Information content (IC) of related concepts may be considered for inclusion in this proposal which could further enhance the effectiveness of retrieval system.

In [52] author's proposal was search model for SW. This model utilized fuzzy formal concept analysis for automatic construction of ontology. Rough set approximations performed the match between query and ontological concepts. Information content (IC) provides useful information regarding search concepts that may be augmented in this proposal for making system more realistic.

In [11] semantic search model using FCA and RST was proposed. This model takes advantage of Wikipedia [10] for concept similarity computation. Proposal may include YAGO [53], WordNet or any other knowledge resource other than Wikipedia. This model overcomes the limitations of existing semantic search models especially models

based on Information content (IC) approach. This model works well for general domains but considered less suitable for specialized domain as Wikipedia does not ensure coverage of specific domains.

5 Discussion and Analysis

This survey presents a comparison of various techniques applied in intelligent IR models and their limitations in the context of SW. A comparative analysis of IR models surveyed is presented in Table 2. In the survey, we observed how different techniques are utilized by these models. The Fig. 2(a) shows that some of preferred techniques in IR for semantic search are query refinement (QR), use of existing corpus (WordNet and Wikipedia), formal concept analysis (FCA), explicit semantic analysis (ESA) and automatic ontology generation. Furthermore, it can be noticed from Fig. 2(b) that 41% surveyed models exploited additional corpus, while 27% models utilized FCA and 32% models used QR, ESA and automatic ontology generation.

Table 2. Comparative analysis of semantic web-based information retrieval models.

S. No.	Models	Techniques	Limitation and challenges
1	Concept-based IR model exploiting term clustering technique and Okapi algorithm [30]	Term clustering, Frequency weighted search, Okapi algorithm, proximity search, Roget's Thesaurus, WordNet	Model may need significant modifications before deploying into more specialized domains like medical as concepts are overlapping in multiple clusters
2	Fuzzy FCA and concept hierarchies-based IR model using domain ontology [31]	Fuzzy FCA, concept hierarchies, automatically built domain ontologies, concept lattice, WordNet	Other fuzzy relations (like resemblance and tolerance) may be employed in fuzzy relational ontological modeling
3	Concept-based IR model with Explicit Semantic Analysis and Wikipedia [28]	Concept extraction, Explicit Semantic Analysis, Wikipedia, feature selection	This model requires existence of knowledge domain such as Wikipedia for concept extraction and feature selection
4	Concept-based semantic search on encrypted cloud data [32]	Semantic searchable encrypted scheme, double score weighting formula, encrypted cloud data, Wikipedia	Performance issue needs further investigations as encrypted search is not very efficient
5	Concept-based search engine with cross-lingual support [8]	Cross-Lingual Information Retrieval System (CLIR), ESA, WordNet	WordNet does not assure the coverage of specialized domain concepts

(*continued*)

Table 2. (*continued*)

S. No.	Models	Techniques	Limitation and challenges
6	Multi-Agent Based IIR framework for SW [4]	Multi-Agent systems, RDF, OWL	Expansion and updating of ontology base are challenging task
7	IR system combining Semantic Web and agent paradigm [33]	Ontology, Multi-Agent systems	Lack of appropriate ontology mapping algorithm for mapping heterogeneous ontology to ontology base
8	Web service discovery method based on multiple agents [35]	Software Agent, OWL, domain ontology	Web service composition can be incorporated in this model for creating new web service by merging & reusing existing web services
9	Agent based educational system on semantic web using domain ontology [36]	Ontology, SW, software agent, least common super-concept (structure taxonomy based) similarity measure	Feature-based similarity measures can also be considered for computing semantic similarity between concept and knowledge resource
10	Search engine based on Web Ontology Language [3]	OWL, Protégé ontology editor, Jena semantic framework Reasoner, ontology analyzer tool	Ranking Algorithm can be augmented for ranking of results set in semantic search
11	Ontology based IR system for sports domain [12]	Ontology, SPARQL, semantic query language	System lacks mechanism to deal with partial match between concept and search terms
12	Semantic Web based intelligent IR system [38]	Ontology, Distributed Hash Table (DHT) for load balancing, range queries and fault tolerance	Response time varies greatly for complex and simple queries
13	SW search based on RDF [5]	Ontology, RDF, Semantic inference	Updating domain ontology whenever new one added is challenging
14	Ontology Web Language and Information Retrieval (OWLIR) framework [6]	SW, markup language: DAML + OIL, AeroText system, DAMLJessKB, Ontology	Model can be extended to include results from partial match between query and concept hierarchy
15	Personalized search engine using query clustering technique and semantic web [40]	Query clustering, web semantics, thesaurus, query expansion	Exploitation of page ranking algorithm methods by some commercial entities to seek people attention throw a challenge

(*continued*)

Table 2. (*continued*)

S. No.	Models	Techniques	Limitation and challenges
16	Enhancing semantic inter-operability in Digital Library by intelligent techniques [39]	Artificial Intelligence, Ontology, RDF, Case-Based Reasoning (CBR), Intelligent Agent	Model can be extended to integrate and use other institutional repositories and digital services. Query refinement can be augmented to extend support for user
17	Semantic IR based on query expansion using domain ontology [9]	SPARQL, Ontology, RDF, WordNet	Model needs significant changes before it can be extended to other domains
18	Fuzzy ontology based IR system for transportation domain [2]	Fuzzy ontology, RDF, RDF query languages: RDQL, RQL, SeRQL	Integration of already existing ontologies of same domain with newly created is challenging task
19	Automated annotation of web services using fuzzy set approach [47]	Extended Vector Space Model (VSM), Fuzzy techniques, WordNet	Extended VSM approach can also be applied to semantic relations such as homonyms and hyponyms
20	Automatic approach for generating fuzzy ontology for Semantic Web [54]	Fuzzy Ontology Generation Framework (FOGA), Fuzzy FFCA, Fuzzy ontology, OWL	Other soft computing techniques can also be integrated for further improving performance and effectiveness of system
21	Fuzzy ontology based intelligent IR system [48]	Fuzzy ontology, Query expansion, WordNet	Fuzzy theory and neural network techniques can be augmented to generate Fuzzy ontology automatically
22	Similarity reasoning in formal concept analysis: From one-to-many valued context [49]	Formal concept analysis, type-2 fuzzy sets, Information content approach	Knowledge contained in hierarchical structure of concept lattice is not included in performing similarity measure
23	Ontology similarity measure combining rough set and concept lattice [51]	Ontology mapping framework, Concept lattice, rough set approximations	Information Content of related concepts are not considered in this proposal
24	Rough set and fuzzy FCA based SW search [52]	RSA, Automatic ontology construction, FFCA	Information content of search terms were not considered
25	Combining FCA, RSA and Wikipedia for Semantic Web search [11]	FCA and RSA based proposal where concept similarity is computed with Wikipedia	Model is not suitable for specialized domain as Wikipedia does not guarantee coverage of specific domains

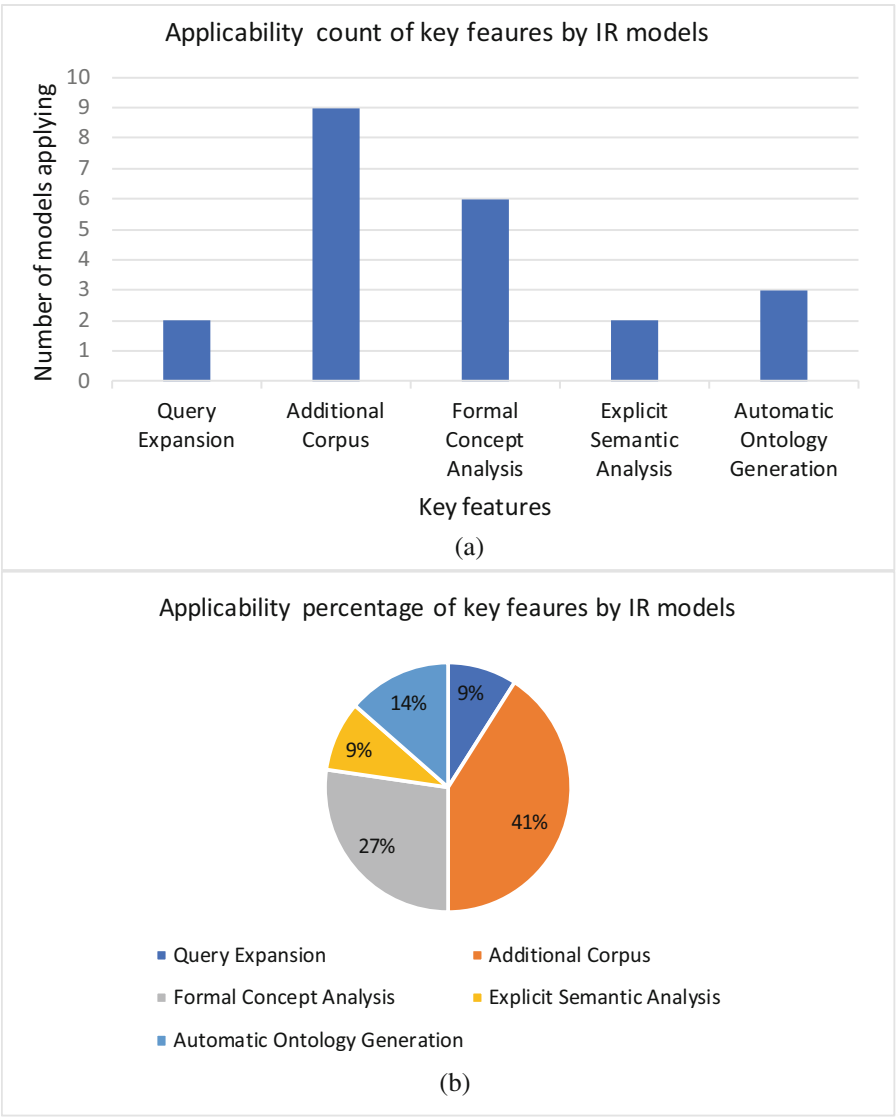

Fig. 2. (a) Applicability count and (b) Applicability percentage of key features by IR models

6 Conclusion and Research Directions

It is evident from literature that traditional IR systems were not effective due to mainly two obstacles. First, they didn't possess methods to deal with the problem of synonyms and polysemy. Second, these methods lack standards for representation, exchange and inference of knowledge encoded in web resources. The use of ontology as knowledge representation tool has overpowered the above problems. Standards like RDF, RDFS,

OWL and DAML + OIL were used to create, model and infer knowledge as domain ontology, which provided basis for intelligent information retrieval systems. This paper presents the significant research work in the field of intelligent information retrieval on SW. Our paper mainly focuses on systems based on multi-agent, ontology, soft computing and concept-based paradigm employed for information retrieval on semantic web. The purpose of this work is to highlight the limitations of these models and techniques employed for IR on SW. The aim of this survey is to focus on the challenges of IR in SW and identification of issues which were not addressed in previous studies of this topic.

IR systems were struggling with issues like dealing with imprecise and vague information, lack of standards for knowledge representation and utilization of semantic knowledge encoded in web resources etc. Researchers found solution of these problems in terms of SW. From this survey, it is observed that researchers are more inclined towards employing SW for IR tasks. Also, survey shows trends of utilization of additional corpus (WordNet, Wikipedia) and formal concept analysis as knowledge processing tool for creating an effective IR system. Application of soft computing techniques for dealing with vagueness in semantic IR systems is also observed in this survey.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
2. Zhai, J., Yan C., Chen, Y.: Semantic information retrieval based on fuzzy ontology for intelligent transportation systems. In: Proceedings of International Conference on Systems, Man and Cybernetics, pp. 2321–2326 (2008)
3. Kumar, S., Singh, M., De, A.: OWL-based ontology indexing and retrieving algorithms for semantic search engine. In: Proceedings of 7th International Conference on Computing and Convergence Technology, pp. 1135–1140 (2012)
4. Xiao, Y., Xiao, M., Zhang, F.: Agents-based intelligent retrieval framework for the semantic web. In: Proceedings of International Conference on Wireless Communications, Networking and Mobile Computing, pp. 5357–5360 (2007)
5. Jiang, J., Wang, Z., Liu, C., Tan, Z., Chen, X., Li, M.: The technology of intelligent information retrieval based on the semantic web. In: Proceedings of 2nd International Conference on Signal Processing Systems, pp. V2-824 (2010)
6. Shah, U., Finin, T., Joshi, A., Cost, R.S., Matfield, J.: Information retrieval on the semantic web. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 461–468 (2002)
7. Miller, G.: WordNet: an on-line lexical database. Int. J. Lexicogr. **3**(4), 235–244 (1990)
8. Samantaray, S.D.: An intelligent concept based search engine with cross linguility support. In: Proceedings of 7th Conference on Industrial Electronics and Applications, pp. 1441–1446 (2012)
9. Chauhan, R., Goudar, R., Sharma, R., Chauhan, A.: Domain ontology based semantic search for efficient information retrieval through automatic query expansion. In: Proceedings of International Conference on Intelligent Systems and Signal Processing, pp. 397–402 (2013)
10. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of 17th conference on Information and Knowledge Management, pp. 509–518 (2008)

11. Jiang, Y., Yang, M.: Semantic search exploiting formal concept analysis, rough sets, and wikipedia. Int. J. Semant. Web Inf. Syst. **14**(3), 99–119 (2018)
12. Zhai, J., Zhou, K.: Semantic retrieval for sports information based on ontology and SPARQL. In: Proceedings of International Conference on Information Science and Management Engineering, pp. 395–398 (2010)
13. Thangaraj, M., Chamundeeswari, M.: A survey of agent-based personalized semantic information retrieval. IJCST **2**(3), 448–498 (2011)
14. Ali, A., Ahmed, I.: Concept-based information retrieval approaches on the web: a brief survey. IJAIR **3**(6), 14–18 (2011)
15. Ahmed, M.W., Ansari, M.A.: A survey: soft computing in intelligent information retrieval systems. In: Proceedings of 12th International Conference on Computational Science and Its Applications, pp. 26–34 (2012)
16. Singh, G., Jain, V.: Information retrieval (IR) through semantic web (SW): an overview. In: Proceedings of CONFLUENCE2012-The Next Generation Information Technology Summit, pp. 23–27 (2012)
17. Sharma, A.: Intelligent information retrieval system: a survey. AEEE **3**(1), 63–70 (2013)
18. Balan, S., Ponmuthuramalingam, P.: A study on world wide web information retrieval and web search techniques. IJIRCCE **4**(3), 3532–3535 (2016)
19. Ezhilarasi, K., Kalavathy, G.M.: Literature survey: analysis on semantic web information retrieval methodologies. In: Proceedings of International Conference for Phoenixes on Emerging Current Trends in Engineering and Management, pp. 93–108 (2018)
20. Fernandez, M., Motta, E.: Semantically enhanced information retrieval: an ontology-based approach. J. Web Semant. **9**(4), 434–452 (2011)
21. Melucci, M.: Boolean Model. In: Liu, L., Ozsu, M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-39940-9_917
22. Melucci, M.: Vector-Space Model. In: Liu, L., Ozsu, M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-39940-9_918
23. Fuhr, N.: Probabilistic models in information retrieval. Comput. J. **35**(3), 243–255 (1992)
24. Jones, K.S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments: Part 2. Inf. Process. Manage. **36**(6), 809–840 (2000)
25. Ali, A., Bari, P., Ahmad, I.: Concept-based information retrieval approaches on the web: a brief survey. IJCIS **3**(6), 14–18 (2011)
26. Woods, W.: Conceptual indexing: a better way to organize knowledge. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories (1997)
27. Ozcan, R., Aslandogan, Y.A.: Concept based information access using ontologies and latent semantic analysis. Technical report CSE-2004-8, University of Texas at Arlington (2004)
28. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. **29**(2), 1–38 (2011)
29. Gottron, T., Anderka, M., Stein, B.: Insights into explicit semantic analysis. In: Proceedings of 20th International Conference on Information and Knowledge Management, pp. 1961–1964 (2011)
30. Henstock, P.V., Pack, D.J., Lee, Y.S., Weinstein, C.J.: Toward an improved concept-based information retrieval system. In: Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval, pp. 384–385 (2001)
31. Goyal, P., Behera, L., McGinnity, T.M.: An information retrieval model based on automatically learnt concept hierarchies. In: Proceedings of International Conference on Semantic Computing, pp. 458–465 (2009)

32. Boucenna, F., Nouali, O., Kechid, S.: Concept-based semantic search over encrypted cloud data. In: Proceedings of 12th International Conference on Web Information Systems and Technologies, pp. 235–242 (2016)
33. Cheng, X., Xie, Y., Yang, T.: Study of multi-agent information retrieval model in semantic web. In: Proceedings of International Workshop on Geoscience and Remote Sensing, pp. 636–639 (2008)
34. Luo, J., Xue, X.: Research on information retrieval system based on semantic web and multi-agent. In: Proceedings of International Conference on Intelligent Computing and Cognitive Informatics, pp. 207–209 (2010)
35. Benaboud, R., Maamri, R., Sahnoun, Z.: Semantic web service discovery based on agents and ontologies. IJIMT **3**(4), 467–472 (2012)
36. Reddy, A.B., Govardhan, A.: A novel approach for similarity and indexing-based ontology for semantic web educational system. IJIEI **4**(2), 117–134 (2016)
37. Yibing, S., Qinglong, M.: Research of literature information retrieval method based on ontology. In: Proceedings of International Conference on Multi-sensor Fusion and Information Integration for Intelligent Systems, pp. 1–6 (2014)
38. Raj, T.F.M., Ravichandran, K.S.: A novel approach for intelligent information retrieval in semantic web using ontology. WASJ **29**, 149–154 (2014)
39. Martín, A., León, C., López, A.: Enhancing semantic interoperability in digital library by applying intelligent techniques. In: Proceedings of SAI Intelligent Systems Conference, pp. 904–911 (2015)
40. Prakasha, S., Shashidhar, H.R., Raju, G.T.: Structured Intelligent Search Engine for effective information retrieval using query clustering technique and semantic web. In: Proceedings of International Conference on Contemporary Computing and Informatics, pp. 688–695 (2014)
41. Rahman, A., Beg, M.M.S.: Face sketch recognition using sketching with words. Int. J. Mach. Learn. Cybern. **6**(4), 597–605 (2014). https://doi.org/10.1007/s13042-014-0256-y
42. Zadeh, L.: Fuzzy Sets. Inf. Control **8**(3), 338–353 (1965)
43. Pawlak, Z.: Rough sets. Int. J. Inf. Comput. Sci. **11**(5), 341–356 (1982). https://doi.org/10.1007/BF01001956
44. Cohen, D., Ai, Q., Croft, W.B.: Adaptability of neural networks on varying granularity IR tasks. In: Proceedings of ACM SIGIR Workshop on Neural Information Retrieval (2016)
45. Chen, H.: Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms. J. Assoc. Inf. Sci. Technol. **46**(3), 194–216 (1995)
46. Cordon, O., Moya, F., Zarco, C.: A new evolutionary algorithm combining simulated annealing and genetic programming for relevance feedback in fuzzy information retrieval systems. Soft. Comput. **6**(5), 308–319 (2002). https://doi.org/10.1007/s00500-002-0184-8
47. Chotipant, S., Hussain, F.K., Hussain, O.K.: An automated and fuzzy approach for semantically annotating services. In: Proceedings of International Conference on Fuzzy Systems, pp. 1–7 (2015)
48. Hourali, M., Montazer, G.A.: An intelligent information retrieval approach based on two degrees of uncertainty fuzzy ontology. Adv. Fuzzy Syst. **2011**, 11 (2011)
49. Formica, A.: Similarity reasoning in formal concept analysis: from one- to-many-valued contexts. Knowl. Based Syst. **60**(2), 715–739 (2019). https://doi.org/10.1007/s10115-018-1252-4
50. Sharma, A., Kumar, S.: Soft computing: dealing with vagueness in intelligent information retrieval. IJRECE **6**(2), 1784–1788 (2018)

51. Zhao, Y., Halang, W.: Rough concept lattice-based ontology similarity measure. In: Proceedings of the First International Conference on Scalable Information Systems (2006)
52. Formica, A.: Semantic web search based on rough sets and fuzzy formal concept analysis. Knowl. Based Syst. **26**, 40–47 (2012)
53. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a large ontology from wikipedia and wordnet. J. Web Semant. **6**(3), 203–217 (2008)
54. Tho, Q.T., Hui, S.C., Cheuk, A., Fong, M., Cao, T.H.: Automatic fuzzy ontology generation for semantic web. IEEE Trans. Knowl. Data Eng. **18**(6), 842–856 (2006)