INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Literature Review: Semantic Web tools and technologies for creating and effective searching method based on the data available in social networking sites

Aamir Junaid Ahmad
aamir.junaid@yahoo.com
Birla Institute of Technology, Mesra, Jharkhand

Sabina Priyadarshini
sabinapriyadarshini@yahoo.co.in
Birla Institute of Technology, Mesra, Jharkhand

ABSTRACT

This paper is a literature review on Semantic Web tools and technologies for creating and effective searching method based on the data available in Social Networking sites. We discuss the requirements of ontologies in the context of the Semantic Web. We survey tools for managing and applying ontologies. Advantages of using ontologies in both knowledge-base-style and database-style applications are compared. We explore the current state of the Semantic Web with a major focus on search based on Social Networking profiles. The paper includes a review of several papers on semantics and ontology-based search, sections on query languages and knowledge base systems that enable Semantic Web search. Finally, we discuss the growth of Social Networking.

Keywords— *Semantic Web, Ontology, SPARQL, OWL, RDF*

1. INTRODUCTION

The main idea of this survey paper is to explore the current state of the Semantic Web (SW) with a major focus on search. The paper includes introductory sections on the SW and its stack, a review of several papers on semantics and ontology-based search, sections on query languages and knowledge base systems that enable SW search. Our survey is by no means complete. There have been a lot of works published on the above topics; we could include only a small portion of them. There exist several other surveys on query languages (e.g. [40] and [44]) and knowledge base systems (e.g., [14], [16] and [44]). We do not follow their style to describe as many languages/tools as possible, but rather tend to give a more detailed overview of used architectures, ideas, etc. For example, our section on knowledge base systems complements those surveys with new entries such as Semagix Freedom, TAP, OWLJessKB, and DLDB, as well as with more recent specifications for Sesame, Jena and KAON.

1.1 Organization.

Sections 2 and 3 introduce the Semantic Web and its stack. Section 4 describes three types of semantics for the Semantic Web. Types of search and ontology based information retrieval model are explained in Sections 5 and 6 respectively. Section 7 provides a brief overview of query languages for the Semantic Web. Section 8 presents a survey on Knowledge base systems for the Semantic Web. Finally, Section 9 concludes the paper.

2. SEMANTIC WEB

In this section, we have introduced some of the tools and technologies that are used in Semantic Web. It focuses on the theories on which this paper is based.

2.1 Semantic Web

"*The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.*"

-Tim Berners-Lee et al. [20]

The World Wide Web (WWW) is a huge library of interlinked documents that are transferred by computers and presented to people. It has grown from hypertext systems, but the difference is that anyone can contribute to it. This also means that the quality of information or even the persistence of documents cannot be generally guaranteed. Current WWW contains a lot of information and knowledge, but machines usually serve only to deliver and present the content of documents describing the knowledge. People have to connect all the sources of relevant information and interpret them themselves.

Semantic web [3] is an effort to enhance current web so that computers can process the information presented on WWW, interpret and connect it, to help humans to find required knowledge. In the same way, as WWW is a huge distributed hypertext system, the semantic web is intended to form a huge distributed knowledge-based system. The focus of the semantic web is to share data instead of documents.

The main idea of the SW, proposed by Tim Berners-Lee, is to enhance existing data on the Web with machine-interpretable metadata to enable better automation, integration, discovery and reuse across the various application.

"*The Semantic Web is a web of data, in some ways like a global database." and "Leaving aside the artificial intelligence problem of training machines to behave like people, the Semantic Web approach instead develops languages for expressing information in a machine process-able form.*"

-Tim Berners-Lee [17]

There is some confusion about the term "Semantic Web":

"*The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a syntax but no semantics.*"

-John Searle [48]

"*Developing XML as a richer version of HTML was generally a good idea. But what botched the Semantic Web is that promoting a universal syntax does nothing to promote semantics. To avoid further confusion, it would be a good idea to rename it the syntactic web.*"

-John Sowa [30]

We do not go into a philosophical discussion about the correctness of the term "Semantic Web", but rather suggest that the usefulness of the SW idea does not depend on the title we use for it. Tim Berbers-Lee has noted that the *semantic* in Semantic Web means *machine processable* [29].

2.2 Ontology

Ontology [4] is the branch of philosophy that seeks to answer the question "what is there?" In computer science, an ontology is a formal conceptualization of a domain. Typically, it specifies the classes of objects that exist, the relationships amongst those classes, the possible relationships amongst instances of the classes, and constraints over those instances. An ontology also defines terms denoting these classes and relationships as well as individual objects. Current web ontology languages, designed to encode information on and for the web, use the eXtensible Markup Language (XML) both for specifying ontologies and also for making assertions about the world using terms defined in ontologies. A semantic web page begins by listing (as URLs) the locations of the ontologies to be used, then goes on to use those ontologies to make assertions about datasets, human beings, items for sale, etc. An agent [5], on coming to such a page, can import the specified ontologies and use that information to understand the semantics of the ensuing assertions. Agent denotes the piece of software that possesses the properties of autonomy, social ability, reactivity, proactivity, temporal continuity, and goal oriented ness. Multi-agent system consists of a number of agents which are capable of interacting with each other. In these systems, the agents are capable to cooperate, coordinate, and negotiate with each other. Various activities in the Semantic Web-based systems are performed by Semantic Web agents.

2.3 Resource Description Framework (RDF)

A core data representation format for the semantic web is the Resource Description Framework (RDF) [6]. RDF is a framework for representing information about resources in a graph form. It was primarily intended for representing metadata about WWW resources, such as the title, author, and modification date of a Web page, but it can be used for storing any other data. It is based on triples *subject-predicate-object* that form a graph of data. All data in the semantic web use RDF as the primary representation language. RDF itself serves as a description of a graph formed by triples. Anyone can define the vocabulary of terms used for a more detailed description. To allow a standardized description of taxonomies and other ontological constructs, an RDF Schema (RDFS) [7] was created together with its formal semantics within RDF. RDFS can be used to describe taxonomies of classes and properties and use them to create lightweight ontologies.

2.4 RDF Query Language SPARQL

The Simple Protocol and RDF Query Language (SPARQL) [8] is a SQL-like language for querying RDF data. For expressing RDF graphs in the matching part of the query, TURTLE syntax is used.
An example of a SELECT query follows.

```
PREFIX foaf:   <http://xmlns.com/foaf/0.1/>
SELECT?name ?emailid
WHERE { ?x foaf:name ?name .
?x foaf:emailid ?email . }
```

The first line defines namespace prefix, the last two lines use the prefix to express an RDF graph to be matched. Identifiers beginning with a question mark? Identify variables. In this query, we are looking for a resource? x participating in triples with predicates foaf: name and foaf:email and want the subjects of these triples.

2.5 The Web Ontology Language OWL

The Web Ontology Language OWL [9] extends RDF and RDFS. Its primary aim is to bring the expressive and reasoning power of description logic to the semantic web. Three species of OWL are defined.

OWL Lite can be used to express taxonomy and simple constraints, such as 0 and 1 cardinality. OWL DL supports maximum expressiveness while retaining computational completeness and decidability. The DL in the name shows that it is intended to support description logic capabilities. OWL full has no expressiveness constraints but also does not guarantee any computational properties. It is formed by the full OWL vocabulary, but does not any impose any syntactic constraints, so that the full syntactic freedom of RDF can be used.

2.6 FOAF

FOAF (an acronym of Friend of a friend) is a machine-readable ontology describing persons, their activities and their relations to other people and objects. Anyone can use FOAF to describe him- or herself. FOAF allows groups of people to describe social networks without the need for a centralized database.

FOAF is a descriptive vocabulary expressed using the Resource Description Framework (RDF) and the Web Ontology Language (OWL). Computers may use these FOAF profiles to find, for example, Assistant Professors in the Computer Science Department between 30 and 40 years of age. This is accomplished by defining relationships between people. Each profile has a unique identifier (such as the person's e-mail addresses, a Jabber ID, or a URI of the homepage or weblog of the person), which is used when defining these relationships.

3. SOCIAL WEB

3.1 What is a Social Web?

The social web [53] is a set of social relations that link people through the World Wide Web. The Social web encompasses how websites and software are designed and developed in order to support and foster social interaction. These online social interactions form the basis of much online activity including online shopping, education, gaming and social networking websites. The social aspect of Web communication has been to facilitate interaction between people with similar tastes. These tastes vary depending on who the target audience is, and what they are looking for. For individuals working in the public relation department, the job is consistently changing and the impact is coming from the social web. The influence, held by the social network is large and ever-changing. As people's activities on the Web and communication increase, information about their social relationships become more available. Social networking sites such as MySpace and Facebook enable people and organizations to contact each other with persistent human-friendly names. Today hundreds of millions of internet users are using thousands of social websites to stay connected with their friends, discover new friends, and to share user-created content, such as photos, videos, social bookmarks, and blogs, even though mobile platform support for cell phones. The social Web is quickly reinventing itself, moving beyond simple web applications that connect individuals to become an entirely new way of life.

3.2 Growth of Social Media

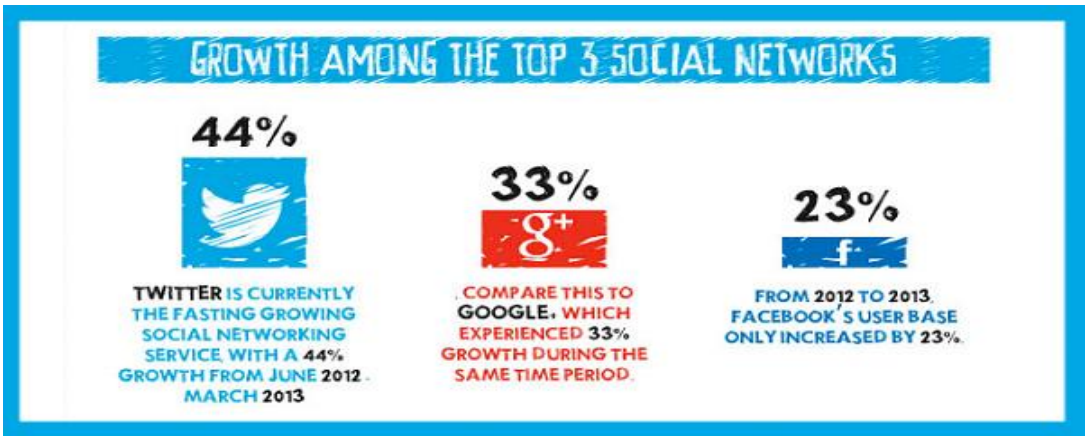


Fig. 1: Growth among top 3 social network

3.3 From Passing Trend to International Obsession

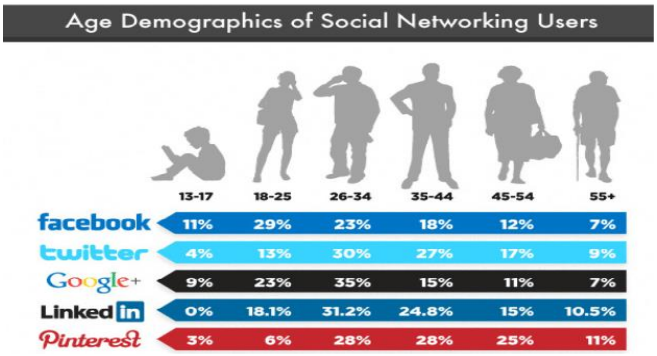
Since 2004, the growth of social media has been near exponential. Back in those days, Facebook — arguably the most mature of the top social networks — only had about 1 million users. By 2011 the network had grown so large, its population was being compared to that of a country. Today, Facebook has more than 1 billion registered users and Mark Zuckerberg has made connecting 5 billion more of a personal mission.

Twitter saw steep growth from 2010 to 2012 but according to the infographic, its growth is starting to slow. Google+ saw the biggest growth of all in 2013, most like because of Google's integration of Google+ into all associated services. If you have a Gmail account, you have a G+ profile. Most recently, Google integrated G+ into YouTube comments. Indeed, Google/Google+ is becoming one big super brand.

While there were many holdouts in the earlier days of social media's development, businesses and marketers love social media now. Indeed, 90 percent of marketers are using social media for business, according to the SEJ infographic. Seventy percent have used Facebook to successfully gain new customers and 34 percent have used Twitter to successfully generate leads.

The infographic [54] includes data from a recent Pew study indicating that 72 percent of all Internet users are also social media users. That means we're well beyond early adopters and social media is becoming as ubiquitous as the computer itself.

2.4 Social Networking Users Age Demographics



3.5 Comparison of Top Networking sites

3.5.1 Number of Users on Popular Social Networking Sites [55]

Table 1: Number of Users on Popular Social Networking Sites

Social Networking Sites	Number of Users(in millions)
Facebook	901
Twitter	555
Google+	170
Linkedin	150
Pinterest	11.7

3.5.2 Unique Monthly Visits on Top Social Networking Websites

Table 2: Unique Monthly Visits on Top Social Networking Websites

Social Networking Sites	Unique Monthly Visitors (in millions)
Facebook	7012.9
Twitter	182.1
Google+	61.0
Linkedin	85.7
Pinterest	104.4

3.5.3 Male-Female Ratio

Table 3: Male-Female Ratio

Social Networking Sites	Male	Female
Facebook	40%	60%
Twitter	43%	57%
Google+	63%	37%
Linkedin	55%	45%
Pinterest	31.8%	68.2%

3.5.4 Time Spend by Average Social networking user per month

Table 4: Time Spend by Average Social networking user per month

Social Networking Site	Time Spend per month (in Minutes)
Facebook	405
Twitter	21
Google+	3
Linkedin	17
Pinterest	89

3.5.5 Estimated User Worth of popular Social networking sites

Table 5: Estimated User Worth of popular Social networking sites

Social Networking Sites	Per User worth
Facebook	$118
Twitter	$71.43
Linkedin	$71
Pinterest	$28.09

4. THE SEMANTICWEB STACK

The SW stack, which describes its components and their relationships, is shown in Figure 1 (copied from [30]).

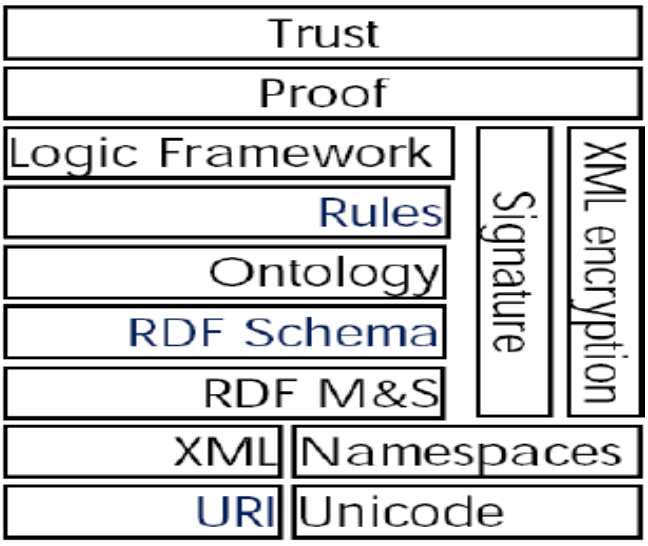


Fig. 2: Semantic web Stack

In the following, we give a brief description of each SW stack layer.

4.1. Layer 1 - URIs and Unicode

Each object on the SW is identified by a unique URI (Uniform Resource Identifier) [19] assigned to it. There are different subclasses of URIs, such as Universal Resource Locators (URLs) and Uniform Resource Name (URNs).

It is important to realize, that the SW will include not only resources like Web pages, images, audio or video files, but also objects like people, events or places. Resources on the Web have unique URLs, however, there is no standard way to assign URIs to people or events. Unicode is a character set that can deal with multiple human languages.

4.2. Layer 2 - XML and Namespaces

The SW metadata uses XML syntax.
Extensible Markup Language (XML) [24] is a standard text format for serializing data using tags. XML has been around for about a decade and has many technologies and tools available for XML data processing, such as DOM and SAX parsers, DTD and XML Schema validation, XPath and XQuery query languages, XML databases, etc.

XML Namespaces [23] are extensions to XML, which provide the mechanism to uniquely identify the element in the vocabulary, where the vocabulary consists of XML element types and attribute names. In multiple XML documents, vocabularies can overlap, leading to problems of recognition and collision. In simple words, the namespace mechanism defines a URI to indicate the vocabulary, and an element name to indicate the element in the vocabulary.

4.3. Layer 3 - RDF

Resource Description Framework (RDF) [15] is a general-purpose language for representing information in the Web. RDF Model is the RDF graph, whose nodes are represented by RDF URI references, blank nodes or plain literals, and arcs are labeled with RDF URI references. (Note that an RDF URI is a URI with restrictions on allowed characters.)

An example of an RDF graph is shown in figure 2 (copied from [30]). The RDF graph consists of triples, where each triple consists of a subject, a predicate and an object. Each triple translates into a statement about a resource. For example, the triple showed in figure 3 (copied from [30]) has the following interpretation: the creator-predicate identified by URI http://purl.org/dc/elements/1.1/creator of the resource (object) identified by URI http://www.example.org/index.html is the object identified by URI http://www.example.org/staffid/85740.

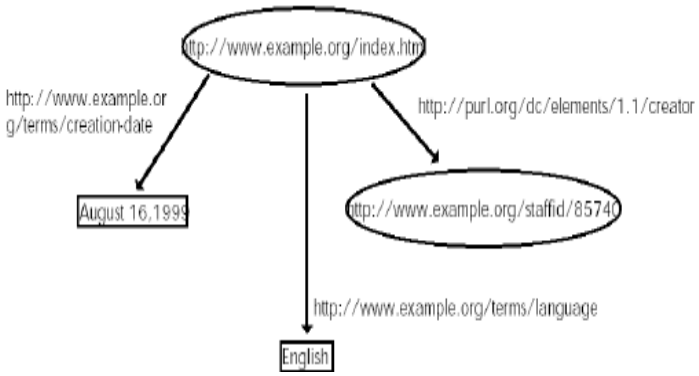


Fig. 3: an RDF graph example

4.4. Layer 4 - RDF Schema

RDF Schema (RDFS) [25] is a language to describe RDF vocabularies (ontologies). RDFS allows describe class and property hierarchies; give labels to URIs; constrain domain and range of properties; etc.

Although RDF Schema and XML Schema [34] are both "schemas", they have different purposes: RDFS is for inference and XMLS is for validation.

4.5 What is Ontology

One of the frequently cited definitions of an ontology is given by Gruber [36]: an ontology is a "formal specification of a conceptualization", and is shared within a specific domain. In other words, an ontology is a document, defined in a formal language (like RDF Schema), that describes a vocabulary of terms or concepts (and their relationships) used for a specific domain.

4.6 Layer 5 - Ontology Languages

Two ontology languages, DAML+OIL [32] and OWL[13], are more complex than RDFS and provide more capabilities to define ontologies. Web Ontology Language (OWL) is the successor of DAML+OIL, therefore we focus our discussion on the former only. OWL extends RDFS with the following features:

property characteristics (TransitiveProperty, SymmetricProperty, FunctionalProperty, inverseOf, InverseFunctionalProperty), property restrictions (all- ValuesFrom, someValuesFrom, cardinality, hasValue), ontology mapping (equivalence between classes and properties – equivalentClass, equivalentProperty; identity between individuals – sameAs; different individuals – differentFrom, AllDifferent), complex class definitions (set operators – intersectionOf, unionOf, complementOf; enumerated classes – oneOf; disjoint classes – disjointWith), etc.

The OWL language provides three increasingly expressive sublanguages: OWL Lite (primarily, a classification hierarchy and simple constraint features), OWL DL (maximum expressiveness with a guarantee of reasoning completeness and decidability) and OWL Full (maximum expressiveness and the syntactic freedom of RDF with no computational guarantees).

4.7. Layer 6 - Rules

Rules are statements that can be used to infer (discover) expressions (knowledge). There are no standards on how to create rules for the SW. The idea of rule statements is not new. For example, in logical programming languages (in particular, Prolog) the rule statement "*male*(*X*): ¡*son*(*X;*)" might mean that an object (atom) is a male if this object is a son of some other object. Prolog uses backward chaining (top-down resolution) for reasoning. Also, a similar idea is used in "deductive databases", where reasoning rules are stated in datalog languages with Prolog-like syntax. The example of a data processor for the SW is the Closed World Machine (cwm) [1]. It is a forward chaining reasoned which can be used for querying, checking, transforming and filtering information. Its core language is RDF, extended to include rules, and it uses RDF/XML or RDF/N3 serializations [1]. The above rule in Prolog can be rewritten using RDF/N3 (N3, Notation 3, is a language to express RDF in a different notation/syntax):

{ ?x a :Male } =>{ ?x :son ?y }

The successful experience of using ontologies and rules on the Web is the Simple HTML Ontology Extensions (SHOE) [10], which uses special HTML tags for ontology definition as well as for creating inference rules. The current version of OWL does not contain special structures to define rules, however some simple "rules" can be designed. For example, Prolog rule

"*male*(*X*) : ¡*man*(*X*)" is similar to OWL definition
<owl:Classrdf:ID="Male">
<owl:equivalentClassrdf:resource="Man"/>
</owl:Class>
which means that classes "Male" and "Man" are equivalent terms or any instance (object, atom) of "Male" is also an instance of "Man" and vice versa. However, for the above Prolog rule, the reasoner, knowing that an object is a male, cannot infer that the very object is a man (unless we also state "*man*(*X*): ¡male(X)").

4.8. The Other Layers

To finish the SW stack discussion, we briefly describe the purpose of the rest layers (Logic Framework, Proof, Trust) and components (Signature, XML encryption).

Logic Framework specifies a formalism for SW reasoning (e.g. Description Logic).

A signature, XML encryption (as well as Proof and Trust) are related to data security issues of the SW. "*As an open and distributed system, the Semantic Web bears the spirit that"anybody can say anything on anybody".*

"People all over the world might assert some statements which can possibly conflict. Hence, one needs to make sure that the original source does make a particular statement (proof) and that source is trustworthy (trust)."

-Shiyong Lu et al. [43]

5. SEMANTICS FOR THE SEMANTICWEB

Sheth et al. [50] described three types of semantics for the SW: the implicit, the formal and the powerful. The implicit semantics is not stated explicitly and extracted from the patterns in data. For example, keyword occurrences, hypertext links, position in

concept hierarchy, etc. This kind of semantics allows finding the relevance of data (document) to some semantic context, however, it is not machine-processable it is not possible to name a relationship between concepts.
The formal semantics is presented in some well-formed syntactic language. The formal language should include the following features:

(a) The notions of model and model-theoretic semantics – language expressions are interpreted in models which reflect "structure of the world", and
(b) The principle of compositionality – expression meaning is a function of the meanings of expression's parts and of the way they are syntactically combined. Examples of such languages are RDF, OWL, and Description Logics. This type of semantics is machine-processable. The major drawback of the formal semantics is that it becomes impractical as knowledge base size increases or knowledge is added from different sources.

The powerful (soft) semantics can exploit implicit and formal semantics (probably "incomplete") to derive relationships using statistical analysis (e.g., probabilistic and fuzzy knowledge). The derived relationships are associated with likelihoods of being valid. The major drawback of the powerful semantics is prior assignments of probabilities to deal with uncertainties.
In summary, the current Web mostly exploits the implicit semantics (search engines like Google), the major focus of the SW is on formal and powerful semantics.

6. TYPES OF SEARCH

Guha et al. [37] identify two kinds of searches:

6.1 Navigational searches

In this class of searches, the user provides the search engine with a phrase or combination of words which s/he expects to find in the documents. There is no straightforward, reasonable interpretation of these words as denoting a concept. In such cases, the user is using the search engine as a navigation tool to navigate to a particular intended document.

6.2 Research searches

In many other cases, the user provides the search engine with a phrase which is intended to denote an object about which the user is trying to gather/research information. There is no particular document which the user knows about that s/he is trying to get to. Rather, the user is trying to locate a number of documents which together will give him/her the information s/he is trying to find.

Example: A search query like "W3C track 2 pm Panel" does not denote any concept. The user is likely just trying to find the page containing all these words. On the other hand, search queries like "Eric Miller" or "Dublin Ohio", denote a person or a place. The user is likely doing a research search on the person or place denoted by the query (copied from [37]).
Both types of searches can be enhanced by exploiting relevant domain ontology and annotations.

7. ONTOLOGY-BASED INFORMATION RETRIEVAL MODEL

In this section, we explore changes that the SW brings to the IR model. The view of the classical IR model is shown in Figure 4 (copied from [35]). In the IR model, a query, formulated from an information need, is matched over document representations (e.g. index structures).

Fig. 4: A classical IR model

Figure 5 (copied from [35]) illustrates the general ontology-based IR model. Information need and queries are refined and formulated based on an ontology structure. Resources (documents) are represented by annotations; ontologies are intended for inference. A matching process is replaced by an exploration process, which can further use an ontology for navigation and logical reasoning. Finally, resource annotations are searched for relevant resources

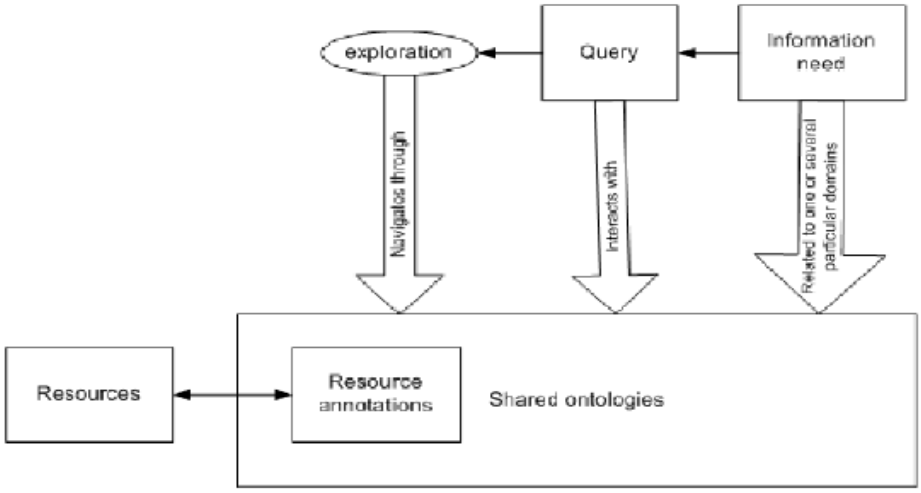


Fig. 5: An ontology based IR model

8. QUERY LANGUAGES FOR THE SEMANTICWEB

In this section, we describe a few RDF(S) query languages as there is extensive research in this area, as well as practical applications in RDF(S) storage and querying systems.

8.1. The Need for RDF Query Languages

Our discussion of the need for RDF query languages is based on [28]. We consider RDF(S) querying on three levels of abstraction:
(a) Syntactic level (XML documents).
(b) Structure level (a set of triples).
(c) Semantic level (graphs with partially predefined semantics).

At the syntactic level, it seems possible to query RDF using XML query languages such as XPath [31] and XQuery [21]. However, the RDF data model is a graph, not a tree, and moreover, both its edges (properties) and its nodes (subjects/objects) are labelled. Relationships in the RDF data model that are not apparent from the XML tree structure become very hard to query. Finally, there are many different ways to serialize the same content using RDF syntax [28].

At the structure level, RDF documents are represented as a set of triples (Subject, Predicate, and Object). The advantage of such a query is that it directly addresses the RDF data model and that it is, therefore, independent of the specific syntax that has been chosen to represent the data. However, a disadvantage of any query language at this level is that it interprets any RDF model only as a set of triples, including those elements which have been given a special semantics in RDFS [28]. For example, information of class and property hierarchies described in RDFS is not exploited in such query languages (e.g., SquishQL [6] and RDQL [7]).

Finally, at the semantic level, the semantics of RDFS descriptions is also used for querying. For example, a query which retrieves all instances of some class will also retrieve instances of that class subclasses. An example of an RDF query language at this level is RQL [41].

8.2 Overview of RDF Query Languages

There is no standard language to query ontologies and annotations. Recently, there have been proposed several RDF query languages, such as RQL [41], SeRQL [26], Versa [46], N3 [18], TRIPLE [12], RDQL [7], SquishQL [6], etc. We give a very brief description of these languages, however, provide a better introduction to RQL.

RQL is an OQL-like typed functional language, which defines a set of basic functions (queries) and iterators and relies on functional composition to build more complex queries. One of the RQL distinguishing features is internal support of schema (RDFS) queries and smooth combination of schema and data querying. Internally, RQL is based on interpretation of the RDF graph.

SeRQL is an attempt to design more powerful and easy to-use querying and transformation language based on existing ideas of RQL, RDQL, N3, etc. Syntactically, it is similar to RQL, however, it is based on interpretation of the RDF Model Theory.
Versa takes an interesting approach in that the main building block of the language is a list of RDF resources. RDF triples play a role in the so-called traversal operations [40].

N3 provides an optional notation (text-based syntax) to express RDF triples, as well as the capability to state rules and queries. N3 does not distinguish between rules and queries. N3 query language is based on the RDF data model.

TRIPLE is derived from F-Logic, such that RDF triples (*Subject; Predicate; Object*) are represented as F-Logic expressions*Subject*[*Predicate¡ > Object*]. TRIPLE is similar to N3, as they both "rule-oriented".

RDQL has a SQL-like syntax similar to RQL and SeRQL. RDQL does not provide support for schema (RDFS) queries. RDQL is based on the RDF graph.

SquishQL is similar to RDQL in its SQL-like syntax, interpretation of data model and capabilities. Research on SquishQL is discontinued.

The comparison of six RDF query languages is presented in [40]. Authors explore expressiveness, closure, adequacy, orthogonality, and safety properties of chosen languages.

Additionally, authors construct 28 queries to evaluate each language: count 0.5 of a point if language succeeds in formulating and evaluating a single query. The resulting ranking (maximum score is 14) is as follows:

- RQL – 10.5
- SeRQL – 8.5
- Versa – 7.5
- N3 – 7.0
- TRIPLE – 5.5
- RDQL – 4.5

More research on standardization and improvement of query languages should be done, as even the leading language, RQL, was not suitable for some query types. Finally, authors [40] include grouping, aggregation, sorting, etc. to the "wish list" of RDF query language features.

In [44], authors similarly provide a detailed theoretical evaluation of expressive power for seven RDF query languages including RQL, SquishQL, Versa, TRIPLE, etc.

9. KNOWLEDGE BASE SYSTEMS FOR THE SEMANTIC WEB

There exist many RDF(S), DAML+OIL and OWL knowledge base systems, which can be used as a foundation for SW repositories. We provide a brief introduction into several such systems, describing their general functionality and architecture, further details on storage design, querying and reasoning support.

9.1. Semagix Freedom

Semagix Freedom [8, 49] is a commercial system which provides support on all levels of SW application development. Semagix Freedom functionality includes:

- Automatic classification of content
- Ontology-driven metadata extraction
- Support for complex query processing involving metadata and ontology
- Ontology design
- Content aggregation
- Knowledge aggregation and creation
- Metadata extraction
- Content tagging and querying of content and knowledge
- Exporting of ontology in RDF/RDFS with some constraints that cannot be expressed in RDF/RDFS

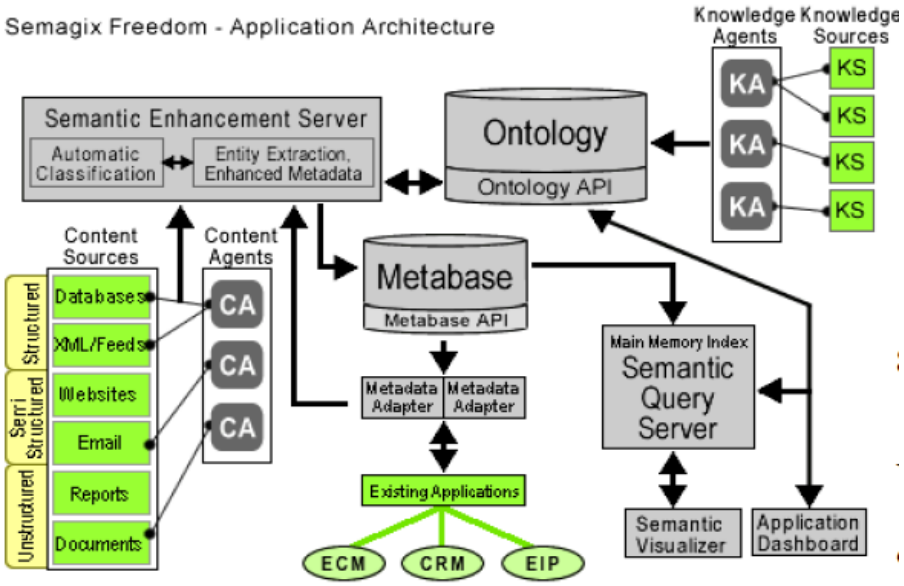


Fig. 6: Semagix Freedom architecture

The system's architecture is presented in figure 6 (copied from [8]). Semagix Freedom provides a modeling tool to design an ontology, Knowledge Agents (KA) automatically maintain the ontology. Metadata is stored in metabase and extracted by Content Agents (CA) from structured, semi-structured and unstructured sources of different formats. Both KA and CA are created without programming using special toolkit based on domain requirements. The extracted content is further "enhanced" by Semantic Enhancement Server: identification of relevant document features (e.g. currencies, dates), entity disambiguation, content annotation (tagging). The metabase is stored in a relational database and its snapshot resides in main memory to facilitate fast querying. Semantic Query Server provides querying mechanism through HTTP and Java APIs, returning results in XML with published DTDs. A few facts about Semagix Freedom performance are listed below (copied from [49]):

- Typical size of an ontology schema for a domain or task ontology: 10s of (entity) classes, 10s of relationships, few hundred property types.
- The average size of the ontology population (number of instances): over a million of named entities.
- A number of instances that can be extracted and stored in a day (before human curation, if needed): up to a million per server per day.
- A number of text documents that can be processed for automatic metadata extraction per server per day: hundreds of thousands to a million.
- Performance for search engine type keyword queries: well over 10 million queries per hour with approx. 10ms per query for 64 concurrent users.
- Query processing requirement observed in an analytical application: approx. 20 complex queries (involving both Ontology and Metabase) to display a page with analysis, taking a total of 1/3 second for computation (roughly equivalent to 50+ query over a relational database with response time over 50 seconds).

9.2 Sesame

Sesame [27, 9, 28] is an RDF framework with support for RDF Schema inferencing. Its main features include querying in three languages (SeRQL, RDQL, RQL), parsing and writing in several serialization syntaxes, support for MySQL, PostgreSQL, Oracle

and SQL Server as well as in memory. It can be deployed as an RDF database, with persistence in an RDBMS, or as a Java library for embedded use in applications [27].

The architecture of Sesame is presented in Figure 7 (copied from [27]). Client programs use Sesame Access APIs to access the server locally or remote through HTTP and RMI. Functional modules like Query Module (SeRQL, RQL, RDQL), Export Module (data exporting into the RDF(S) format) and Admin Module (administrative functionality) are clients of SAIL API – Storage And Interface Layer. SAIL layer is the main component of Sesame that provides an application programming interface that abstracts from the storage device used (in-memory storage, disk-based storage, RDBMS) and takes care of inferencing. Note, that queries in SeRQL, RQL and RDQL are translated into a sequence of SAIL API calls. Therefore, a substantial part of the query evaluation process is done in the respective query modules [28]. As already stated above, Sesame can store RDF data into a relational database (MySQL, Oracle and SQL Server) or into an object-relational database (PostgreSQL). We briefly describe the relational schema for these two storage approaches in the following (details are available in [28]).

Sesame and MySQL the relational schema (see Figure 8) is fixed and includes tables to store RDFS ontology (*class, subClassOf, property, subPropertyOf, domain, range, type, labels*, etc.) and single table *TRIPLES* for RDF statements.

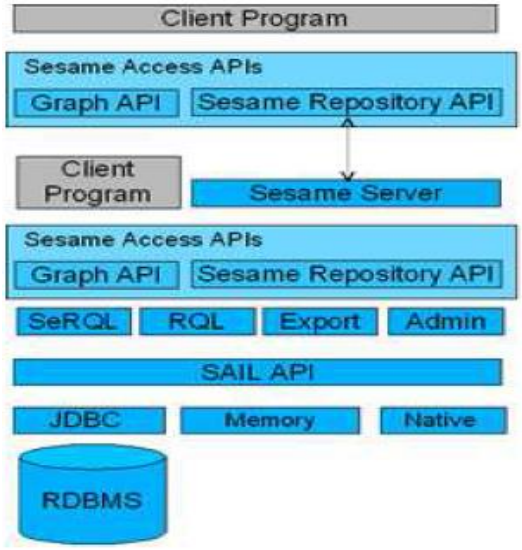


Fig. 7: Sesame Architecture

9.3 Jena

Jena [2, 52, 51] is a Java framework for building Semantic Web applications, that provides a programmatic environment for RDF, RDFS and OWL, including a rule-based inference engine. Jena stores RDF graphs in memory or in a database and supports two forms of querying: triple match and RDQL [7]. In the following, we briefly describe persistent RDF storage, querying and inferencing in Jena.

Persistent RDF storage has evolved from Jena1 to Jena2, the second generation of the Jena toolkit. Jena1 used the normalized triple store approach (similar to Sesame) illustrated in Figure 8 (copied from [51]). The statement table stores all triples (statements) and references resources and literals tables for subjects, predicates and objects. The normalized schema in Jena1 was used for MySQL, PostgreSQL and Oracle, however, Berkley DB used denormalized schema storing triples in a single table. This approach is very efficient in space, however, requires a three-way join to retrieve a triple.

Jena2 exploits the denormalized schema approach (see Figure 9 (copied from [51])). Separate literals and resources tables are only used to store values of long literals and URIs respectively when their length exceeds some threshold (can be specified). This approach takes more space (literal and URI values may be stored repeatedly), however, shows better retrieval performance.

Statement Table

Subject	Predicate	ObjectURI	ObjectLiteral
201	202	null	101
201	203	204	null
201	205	101	null

Literals Table

Id	Value
101	Jena2
101	The description - a very long literal that might be stored as a blob.

Resources Table

Id	URI
201	mylib:doc
202	dc:title
203	dc:creator
204	hp:JenaTeam
205	dc:description

Fig. 8: Jena1 Normalized Schema

Statement Table

Subject	Predicate	Object
mylib:doc1	dc:title	Jena2
mylib:doc1	dc:creator	HP Labs - Bristol
mylib:doc1	dc:creator	Hewlett-Packard
mylib:doc1	dc:description	101
201	dc:title	Jena2 Persistence
201	dc:publisher	com.hp/HPLaboratories

Literals Table

Id	Value
101	The description - a very long literal that might be stored as a blob.

Resources Table

Id	URI
201	hp:aResource-WithAnExtreme-lyLongURI

Fig. 9: Jena2 Denormalized Schema

10. CONCLUSION

In this survey, we reviewed papers related to Semantic Web search: semantics for search, ontology-based search, query languages and knowledge base systems that enable semantic web search. The paper complements existing surveys with new systems, more detailed and recent specifications on some of the systems.

Our conclusions drawn from this survey include:

(a) Existing RDF(S) query languages are not complete, lacking expressivity.

(b) Lack of standards on query languages, rules and inferencing resulted in many different incompatible implementations, which makes them difficult to compare, learn, exploit, and so forth. It is not always clear which inferencing path a reasoner should choose as correct.

(c) Existing systems are not mature and have the following limitations:
- Performance and scalability, which are significantly influenced by inefficient storage schemas and inference algorithms,
- Low expressivity of implemented query languages,
- Completeness of query results, which is influenced by insufficient support of inference,
- The soundness of query results, which is influenced by incorrect inference,
- Lack of performance/ scalability experiments.

11. REFERENCES

[1] The cwm home page. http://www.w3.org/2000/10/swap/.
[2] Jena – a Semantic Web framework for Java. http://jena.sourceforge.net/.
[3] KAON project. http://kaon.semanticweb.org/.
[4] OWLJessKB: A Semantic Web reasoning tool. http://edge.cs.drexel.edu/assemblies/software/owljesskb/.
[5] The RDF Query Language (RQL). http://139.91.183.30:9090/RDF/RQL/.
[6] RDF query using SquishQL. http://swordfish.rdfweb.org/rdfquery/.
[7] RDQL - RDF Data Query Language. http://www.hpl.hp.com/semweb/rdql.htm.
[8] The Semagix homepage. http://www.semagix.com/.
[9] The Sesame homepage. http://www.openrdf.org/.
[10] The SHOE home page. http://www.cs.umd.edu/projects/plus/SHOE/.
[11] The TAP homepage. http://tap.stanford.edu/.
[12] TRIPLE homepage. http://www.dfki.uni-kl.de/frodo/triple/.
[13] S. Bechhofer, F. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language Reference, February 2004. http://www.w3.org/TR/2004/REC-owl-ref-20040210/.
[14] D. Beckett. SWAD-Europe Deliverable 10.1: Scalability and Storage: Survey of Free software/ Open Source RDF storage systems. http://www.w3.org/2001/sw/Europe/reports/
RDF scalable storage report/.
[15] D. Beckett. RDF/XML Syntax Specification (Revised), February 2014. http://www.w3.org/TR/2004/REC-rdf-syntaxgrammar- 20040210/.
[16] D. Beckett and J. Grant. SWAD-Europe Deliverable 10.2: Mapping Semantic Web Data with RDBMSes. http://www.w3.org/2001/sw/Europe/reports/ scalable RDBMS mapping report/.
[17] [T. Berners-Lee. The semantic Web road map, September 1998.
http://www.w3.org/DesignIssues/Semantic.html.
[18] T. Berners-Lee. Notation 3. 1998. http://www.w3.org/DesignIssues/Notation3.html.
[19] T. Berners-Lee, R. Fielding, and L. Masinter.RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax, August 1998.
[20] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web.Scientific American, May 2001.
[21] [S. Boag, D. Chamberlin, M. F. Fernandez, D. Florescu, J. Robie, and J. Simeon. XQuery 1.0: An XML Query Language, February 2005. http://www.w3.org/TR/2005/WD-xquery-
20050211/.
[22] E. Bozsak, M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, G. Stumme, Y. Sure, J. Tane, R. Volz, and V. Zacharias. KAON - towards a large scale Semantic Web. In Proceedings of the Third International Conference on E-Commerce and Web Technologies, September 2012.

[23] T. Bray, D. Hollander, and A. Layman.Namespaces in XML, January 1999. http://www.w3.org/TR/1999/REC-xmlnames-19990114/.
[24] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau. Extensible Markup Language (XML) 1.0 (Third Edition), February 2004. http://www.w3.org/TR/2004/RECxml- 20040204/.
[25] D. Brickley and R. Guha. RDF Vocabulary Description Language 1.0: RDF Schema,February 2004. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.
[26] J. Broekstra and A. Kampman.SeRQL: An RDF query and transformation language.DRAFT. http://www.cs.vu.nl/ jbroeks/papers/SeRQL.pdf.
[27] J. Broekstra and A. Kampman. RDF(S) manipulation, storage and querying using Sesame. In International Semantic Web Conference (ISWC), demo paper, 2004.
[28] J. Broekstra, A. Kampman, and F. van Harmelen. Sesame: A generic architecture for storing and querying RDF and RDF Schema. In I. Horrocks and J. Hendler, editors, Proceedings of the First Internation Semantic Web Conference, number 2342 in Lecture Notes in Computer Science, pages 54–68.Springer Verlag, July 2002.
[29] M. Butler. Barriers to real-world adoption of semantic web technologies. http://www.hpl.hp.com/personal/marbut/ barriersToRealWorldAdoptRDF.pdf.
[30] M. H. Butler. Is the Semantic Web hype? http://www.hpl.hp.com/personal/marbut/ is TheSemanticWebHype.pdf.
[31] J. Clark and S. DeRose. XML Path Language (XPath) Version 1.0, November 2009. http://www.w3.org/TR/1999/REC-xpath-19991116.
[32] D. Connolly, F. Harmelen, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. DAML+OIL (March 2001) Reference Description, December 2010. http://www.w3.org/TR/2001/NOTE-daml+oil-reference- 20011218.
[33] L. Ding, K.Wilkinson, C. Sayers, and H. Kuno.Application-specific schema design for storing large RDF datasets. In 1st International Workshop on Semantic Web and Databases (SWDB'03, with VLDB03), 2003.
[34] [34] D. C. Fallside and P. Walmsley. XML Schema Part 0: Primer Second Edition, October 2004. http://www.w3.org/TR/2004/REC-xmlschema-0- 20041028/.
[35] E. Garcia and M.-A.Sicilia. User interface tactics in ontology-based information seeking. PsychNology Journal, 1(3):242–255, July 2013.
[36] T. Gruber. A translation approach to portable ontology specifications. Knowledge acquisition, 6(2):199–221, 1993.
[37] R. Guha, R. McCool, and E. Miller.Semantic Search. In The Twelfth International World Wide Web Conference, May 2003.
[38] Y. Guo, Z. Pan, and J. Heflin. Choosing the best knowledge base system for large semantic web applications. In Thirteenth International World Wide Web Conference (WWW2004), pages 302–303, 2004.
[39] Y. Guo, Z. Pan, and J. Heflin. An evaluation of knowledge base systems for large OWL datasets. In Third International Semantic Web Conference, Hiroshima, Japan, pages 274–288, 2004.
[40] P. Haase, J. Broekstra, A. Eberhart, and R. Volz. A comparison of RDF query languages. In The Semantic Web - ISWC2004. Proceedings of the Third International Semantic Web Conference, 2004.
[41] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis, and M. Scholl. RQL: A declarative query language for RDF. The Eleventh International World Wide Web Conference (WWW'02), 2002.
[42] J. B. KopenaandW. C. Regli. DAMLJessKB: A tool for reasoning with the Semantic Web. In 2nd International Semantic Web Conference (ISWC2003), 2003.
[43] S. Lu, M. Dong, and F. Fotouhi. The Semantic Web: Opportunities and challenges for next-generation Web applications. International Journal of Information Research, 7(4), July 2002.
[44] A. Magkanaraki, G. Karvounarakis, T. T. Anh, V. Christophides, and D. Plexousakis. Ontology storage and querying. Technical Report No 308. April 2012.http://139.91.183.30:9090/RDF/publications/tr308.pdf.
[45] B. Motik, D. Oberle, S. Staab, R. Studer, and R. Volz. KAON server architecture. Technical Report 421, University of Karlsruhe, Institute AIFB, 76128 Karlsruhe, Germany. 2013. http://wonderweb.man.ac.uk/deliverables/documents/D5.pdf.
[46] M. Olson and U. Ogbuji.Versa. http://uche.ogbuji.net/tech/rdf/versa/.
[47] Z. Pan and J. Heflin. DLDB: Extending relational databases to support Semantic Web queries. In Workshop on Practical andScaleable Semantic Web Systems, ISWC pages 109–113,2003.1 KIF Knowledge Interchange Format – http://logic.stanford.edu/kif/kif.html
[48] J. R. Searle. Minds, brains, and programs.The Behavioral and Brain Sciences, 3, 1980. Also available at http://www.w3.org/DesignIssues/Semantic.html.
[49] A. Sheth and C. Ramakrishnan. Semantic (Web) technology in action: Ontology-driven information systems for search, integration and analysis. In IEEE Data Engineering Bulletin, Special issue on Making the Semantic Web Real pages 40–48, December 2003.
[50] A. Sheth, C. Ramakrishnan, and C. Thomas. Semantics for the Semantic Web: The implicit, the formal and the powerful. International Journal on Semantic Web and Information Systems, 1(1):1–18, January-March 2013.
[51] K. Wilkinson, C. Sayers, H. Kuno, and D. Reynolds. Efficient RDF storage and retrieval in Jena2. In 1st International Workshop on Semantic Web and Databases (SWDB'03, with VLDB03), pages 131–151, 2003.
[52] K. Wilkinson, C. Sayers, H. A. Kuno, D. Reynolds, and L. Ding. Supporting scalable, persistent Semantic Web applications. IEEE Data Engineering Bulletin, 26(4):33–39, 2003.
[53] Halpin, Harry; Tuffield, Mischa. "A Standards-based, Open and Privacy-aware Social Web". W3C Social Web Incubator Group Report 6th December 2010 Report.
[54] http://socialtimes.com/the-growth-of-social-media-from-trend-to-obsession-infographic_b141318
[55] http://www.go-gulf.com/blog/social-networking-user/

