Chapter 1

Introduction

1.1 What is Web Intelligence?

Today, the world is experiencing the excitement of an historic change. We find ourselves in the midst of an information revolution, the result of rapid advances in technology built in great part upon the shoulders of three pivotal pioneers: Kurt Godel, Alan Turing and Tim Berners-Lee. Through their contributions, we are witnessing the remarkable refashioning of the Information Age, which began in the 1950s, into the Information Revolution as the World Wide Web evolves into a resource with intelligent capabilities.

While the capabilities and scope of today’s World Wide Web are impressive, its continuing evolution into a resource with intelligent features and capabilities presents many challenges. The traditional approach of building information systems has consisted of custom made, costly database applications. However, this is changing. Information services are beginning to use generic components and open global standards to offer widely accessible graphical presentations with easier interaction. As a result, benefits are accruing to transactions over the Web including such areas as: e-commerce, banking, manufacturing and education.
At the heart of the Information Revolution is the transformation of the world toward a knowledge economy with a knowledge society. Helping to forge this transformation is the World Wide Web Consortium (W3C), which is working to deliver global machine processing built upon layers of open markup languages.

It is widely accepted that the technology of today’s Information Age has had a major impact on global communications and commerce and that it will continue to support major improvements in human productivity. However, while the World Wide Web is making significant contributions to this progress, there remain many challenges to its further development into a resource with intelligent features.

For the Information Age to achieve its full potential in improving human productivity, at least two key new advances must still be achieved: (1) ubiquitous access to transaction applications of all types and (2) intelligent software applications enabling automated transactions. For example, Web Services require human processing to be implemented. In addition, Web Services rely on the interoperation of two competing proprietary server frameworks to successfully communicate complex business logic. The solution of the W3C to both of these problems is to deliver automatic machine processing globally through a Web architecture utilizing layers of open markup languages.

The term “intelligence” can be applied to nonhuman entities as we do in the field of Artificial Intelligence (AI). But frequently we mean something somewhat different than in the case of human intelligence. It is recognized that human thinking involves complicated interactions
within the biological components of the brain and that the process of learning is also an important element of human intelligence. Increasingly, software applications perform tasks that are sufficiently complex and human like that the term intelligent may be appropriate. Whereas AI can be seen as the science of machines that behave intelligently (or simulate intelligent behavior), the concept of intelligent applications entails the efforts to take advantage of AI technologies to enhance applications and make them act in more intelligent ways. This brings us to the question of Web intelligence or intelligent software applications on the Web.

The World Wide Web can be described as an interconnected network of networks, but that does not go quite far enough. The present day Web consists not only of the interconnected networks, servers and clients, but also the multimedia hypertext representation of vast quantities of information distributed over an immense global collection of electronic devices. With software services being provided over the Web, one can readily see an analogy to the human (or machine) thinking process where information is stored, accessed, transferred and processed by electronic patterns in electrical devices and their interconnections.

However, the current Web consists primarily of static data representations that are designed for direct human access and use. We can view effective web searches as an information retrieval problem. Search engines are one Web technology designed to
automatically process information from large numbers of Web sites to deliver useful processed information, but the search methods used today have rudimentary capabilities. Between the classical models of IR, the vector space model has been shown experimentally to have better performance over the earlier Boolean method. Many recent IR systems are built on the popular vector space model (VSM). Thus, the search engines based on classical vector space model are syntactic based and look for the partial/full match of the query terms in the documents. This feature being a big success in the IR community suffers with the inability to retrieve the semantically related documents just because they do not contain the keyword.

The key to moving to the next level is the improvement of the ability of software applications to communicate directly with one another and the representation of information in ways that are far more usable by software applications. An important framework for creating such meaningful abilities can be provided by the proposed next generation of Web architecture: the Semantic Web.

### 1.1.1 World Wide Web

How is the World Wide Web managing knowledge and empowering the Information Revolution? Does rapid change and improved information productivity require more intelligent Web capabilities? What technologies offer the best opportunities for sustained powerful
change? This section explores these questions by briefly describing the development and limitations of today’s Web technology.

The first implementation of the web represents the Web 1.0, which, according to Berners-Lee [1], could be considered the "read-only web". In other words, the early web allowed us to search for information and read it. There was very little in the way of user interaction or content contribution. However, this is exactly what most website owners wanted. Their goal for a website was to establish an online presence and make their information available to anyone at any time. Shopping cart applications, which most e-commerce website owners employ in some shape or form, basically fall under the category of Web 1.0.

Currently, we are seeing the infancy of the Web 2.0 or the "read-write" web if we stick to Berners-Lee's [1] method of describing it. The ability to contribute content and interact with other web users has dramatically changed the landscape of the web in a short time. For example, looking at YouTube and MySpace, which rely on user submissions and the potential, becomes clearer. The Web 2.0 appears to be a welcome response to a demand by web users that they be more involved in what information is available to them.

Now, it’s important to realize that there are a staggering number of definitions of what constitutes a "Web 2.0 application". For example, the perception exists that just because a website is built using a certain technology (like Ruby on Rails), or because it employs Ajax in its interface, it is a Web 2.0 application. From the general, bird’s-eye view we are taking, this is not the case; our definition simply requires
that users be able to interact with one another or contribute content. Developers, for example, have a much more rigid definition of Web 2.0 than average web users, and this can lead to confusion.

This in turn leads us to the rumblings and mumblings we have begun to hear about Web 3.0, by extending Tim Berners-Lee's explanations, the Web 3.0 would be something akin to a "read-write-execute" web. With Web 3.0 you'll be able to sit back and let the Internet do all the work for you. You could use a search service and narrow the parameters of your search. The browser program then gathers, analyzes and presents the data to you in a way that makes comparison a snap. It can do this because Web 3.0 will be able to understand information on the Web.

A Semantic Web (or Web 3.0) agent could be programmed to do almost anything, from automatically booking your next vacation to researching a term paper. If you visit a movie blog, for instance, and read about a particular film, it immediately links to sites where you can buy or rent that film.

With the Semantic Web, computers will scan and interpret information on Web pages using software agents. These software agents will be programs that crawl through the Web, searching for relevant information. They'll be able to do that because the Semantic Web will have collections of information called ontologies. In terms of the Internet, ontology is a file that defines the relationships among a group of terms.
For the Semantic Web to be effective, ontologies have to be detailed and comprehensive. In Berners-Lee's concept, they would exist in the form of metadata. Metadata is information included in the code for Web pages that is invisible to humans, but readable by computers.

1.2 Metadata

The rapid increase in the number and variety of resources on the World Wide Web has made inappropriateness of traditional schemas of resource description for web resources and has encouraged significant activities recently on defining web compatible schemas named "metadata".

Metadata, in general, is defined as 'data about data' or 'information about information'. In the other words, metadata is data that describe information resources. Metadata is data that provide extra information about other data. For example, a photo can be described using the following metadata: <dateTaken> 01/01/2011 </dateTaken>, <placeTaken> seminar room </placeTaken> and <whatAbout> project meeting </whatAbout>.

The information being described by metadata, may be considered at the first look as corporal and digital information resources such as books, newspapers, journals, photographs and so on. Greenberg [3] refers to this data as "object" and states that this object is any entity, form or mode for which contextual data can be recorded. The universe of objects to which metadata can be applied is radically diverse and seemingly endless, ranging from corporeal and digital information
resources, such as a monograph, newspaper or photograph, to activities, events, persons, places, structures, transactions, relationships, execution directions and programmatic applications.

Metadata, therefore, captures the wide range of intrinsic or extrinsic information about a variety of objects. These intrinsic or extrinsic characteristics and features are described in the individually structured data elements that facilitate object use, identification and discovery.

The way that current service oriented infrastructure handles and manages services metadata is not adequate and effective for metadata to help services discovery and knowledge sharing. There are no problems for humans to understand XML based metadata because we know the meaning of these English words, the question is: “can machines understand and consume them?”, so that they can perform automatic processing with regards to the use of Web/Grid services. Clearly without further assumptions, the answer will be no. The Semantic Web / Grid are extensions of the current Web/Grid in which information and services are given well defined meaning, better enabling computers and people to work in cooperation. We believe that the first step towards the Semantic Web/Grid is to make the Web/Grid full of rich SMD, in other words, metadata with semantics.

1.3 Ontologies

Ontology is defined as an explicit specification of a conceptualization [4]. An ontology is a formal explicit description of
concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions)). Ontology together with a set of individual instances of classes constitutes a knowledge base.

If a program wants to compare conceptual information across two knowledge bases on the Web, it has to know when any two given terms are being used to mean the same thing. Ideally, the program must have a way to discover common meanings for whatever knowledge bases it encounters. A solution to this problem is provided by the Semantic Web in the form of collections of information called ontologies. A typical ontology for the Web uses taxonomy and a set of inference rules. The taxonomy defines classes of objects and relations among them. Classes, subclasses, and relations among entities are important tools. We can express a large number of relations among entities by assigning properties to classes and allowing subclasses to inherit such properties. Inference rules in ontologies may express rules for manipulating information. Inference rules may express the rule: “If a city code is associated with a state code, and an address uses that city code, then that address has the associated state code”.

Being the conceptual models that capture domain knowledge, ontologies can be looked upon for aiding meaningful information retrieval. Ontology in general contains a vocabulary of terms that refer
to the things of the interest in a given domain, some specification of meaning for the terms grounded in some form of logic.

The way the knowledge of the domain is captured in Ontology enables one to retrieve the related information for a given term, thus supporting intelligent information retrieval. The various languages used for representing the ontologies are RDF/RDFS, DAML+OIL, OWL. The latest W3c recommended standard for representing ontologies is OWL. OWL contains most of the features to express the semantics of the domain knowledge.

Now that we have the syntactic search engines and the knowledge base capturing the semantics of a domain, query expansion mechanism serves as a medium between these two techniques and thus helps in retrieving the semantically relevant information.

1.4 Query Expansion

Logic reasoning is the formal semantic reasoning based on the explicitly defined relations between concepts in the ontology. The main logic reasoning expansions are expansion to equivalent concepts, expansion to broader or narrower concepts and expansion with the concepts having common super class.

Current practice, for example most search engines work at lexical level, retrieving only the documents containing the words from the query. The words from the query do not consist in the relevant documents and is called imprecise retrieval. Query expansion addresses imprecise retrieval by modifying the query, adding in words
related to the original query words. The additional terms may be taken from a thesaurus or calculated in a statistical or probabilistic way. Query Expansion is useful because imprecise queries cannot retrieve the entire set of relevant documents. The intent is to improve precision and/or recall. From the mere words in a query, we cannot tell the exact meaning that a searcher has in mind. We do have resources that give us more information than the query words provide alone. On the other hand, we can aid the query expansion with the domain knowledge to avoid retrieving those documents which are irrelevant even in the presence of the query.

1.5 Inference Engines

An inference engine controls overall execution of a set of rules. It searches through a knowledge base, attempting to pattern-match facts or knowledge present in memory to the antecedents of rules. If a rule’s antecedent is satisfied, the rule is ready to fire and is placed in the agenda. When a rule is ready to fire it means that since the antecedent is satisfied, the consequent can be executed. They deduce new knowledge from previously established knowledge. It requires everyone to share exactly the same definition of common concepts. But central control is stifling, and increasing the size produces complexity that rapidly becomes unmanageable. These systems limit the questions that can be asked reliably. In avoiding the problems, traditional knowledge–representation systems narrow their focus and use a limited set of rules for making inferences. So the proposed
approach is Plausible Inference. Plausible in a real world means having an appearance of truth or reason, means it may not be true but it is still believable. Given the statement that the spare tire is contained in the trunk and the trunk is part of the car, a plausible inference is that the spare tire is contained in the car.

1.6 Description of the Research Work

1.6.1 Motivation

Search engines today are based on decades old technology patched with new solutions. When specifying a search, users enter a small number of terms in the query. Yet the query describes the information need, and is commonly based on the words that people expect to occur in the types of document they seek. This gives rise to a fundamental problem, in that not all documents will use the same words to refer to the same concept. Therefore, not all the documents that discuss the concept will be retrieved by a simple keyword-based search. Furthermore, query terms may of course have multiple meanings (query term polysemy). As conventional search engines cannot interpret the sense of the user's search, the ambiguity of the query leads to the retrieval of irrelevant information.

Converse to the problem of polysemy, is the fact that conventional search engines that match query terms against a keyword-based index will fail to match relevant information when the keywords used in the query are different from those used in the index,
despite having the same meaning (index term synonymy). Although this problem can be overcome to some extent through thesaurus-based expansion of the query, the resultant increased level of document recall may result in the search engine returning too many results for the user to be able to process realistically. In addition to an inability to handle synonymy and polysemy, conventional search engines are unaware of any other semantic links between concepts.

Many search engines fail to take into consideration aspects of the user's context to help disambiguate their queries. User context would include information such as a person's role, department, experience, interests, project work etc.

The results returned from a conventional search engine are usually presented to the user as a simple ranked list. The sheer number of results returned from a basic keyword search means that results navigation can be difficult and time consuming. Generally, the user has to make a decision on whether to view the target page based upon information contained in a brief result fragment.

Most of them use the inverted index method and its statistics/popularity flavored derivatives. The search problem is directly attributable to the limitations of the inverted index method as the underlying platform. Any kind of semantic enrichment requires handling and organizing semantically rich data, and a very short processing time. This requirement exceeds what is expected from an inverted index regardless of hardware capacity.
The search community has provided various solutions to such a problem by means of query expansion and semantic annotation. Query Expansion is addressed using linguistic knowledge in the form of WordNet/thesauri. Query expansion is also addressed using the domain specific knowledge captured in taxonomic model of Ontologies. By the very complex nature of the knowledge representation of a domain, very few relationships are addressed while solving this problem in QE. State of the art query expansion uses synonyms and hierarchical relationships only, expanding either in the directions of parent or children. Semantic annotation requires the annotation of the resources by using the terms from a knowledge base. Such a process of manual creation of the domain knowledge and also annotating the documents with the knowledge is quite expensive. As a result, annotations often are incomplete or erroneous, resulting in decreased search performance.

1.6.2 Proposed Work

Our approach targets towards providing the relevant results using the domain specific knowledge. To meet the expectations/provide the solutions for the problem discussed above, we have used domain knowledge represented in the form of OWL Ontology. Our approach uses query expansion method to expand the query and the traditional basic keyword search as the search mechanism. Our system fits the query terms in the ontology graph in an appropriate way and exploits the surrounding knowledge to
retrieve the relevant results. We find that we can fix the context of the query and also bring in semantically related terms into the picture using this approach. The resulting enhanced query is given to the underlying basic keyword search system. As a result, we find that we can achieve substantial improvement in both precision and recall compared to the basic keyword search system.

We also proposed our approach towards the next generation service-oriented computing infrastructure with rich metadata and semantic support and also presented an integrated framework for Semantic Meta Data management for Web/Grid services.

Finally a case study, describes the applicability of social network analysis to the semantic web, particularly discussing the multi-dimensional networks that evolve from ontological trust specifications.

1.6.3 Issues

It is the nature of the relevance that there is no absolute right and wrong. What is a relevant document for one person’s query might be an irrelevant document for the same query of another person. It is so too with ontologies there are many possible correct relationships between concepts depending on information need.
1.6.4 Scopes and Objectives

The proposed system retrieves relevant documents for a search with an increase in the ratio of precision and recall.

- The objective of the proposed system is to provide a knowledge representation of linked data in order to allow machine processing on a global scale.
- The proposed system holds good for the text corpus.
- The query words map to the vocabulary in the Ontology.
- The scope of the proposed system is limited to a particular domain, for which the knowledge is represented in a very rich structure called ontology.
- The order of the query words in which appear does not matter because they merely map to the terms in the Ontology.

1.6.5 Organization of the thesis

This thesis is organized as follows:

Chapter 2 discusses the literature survey in the direction of the problem statement.

Chapter 3 discusses the new directions in building seamless next generation web.

Chapter 4 discusses the approaches towards next generation service oriented computing infrastructure with rich metadata and semantic support.

Chapter 5 discusses the case study, which describes the applicability of social network analysis to the semantic web,
particularly discussing the multi-dimensional networks that evolve from ontological trust specifications.

Chapter 6 discusses conclusion.