Searching The Semantic Web

Lecturer:
Kyumars Sheykh Esmaili

Semantic Web Research Laboratory
Computer Engineering Department
Sharif University of Technology

Fall 2005
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
Before and After ?

a) Current Web

b) Semantic Web
Semantic Web Terminology

- A **Term** is a non-anonymous RDF resource which is the URI reference of either a **class** or a **property**.

- An **Individual** refers to a non-anonymous RDF resource which is the URI reference of a class member.

- An **Ontology** contains mostly term definition (i.e. classes and properties). It corresponds to T-Box in Description Logic.

- An **Annotation** contains mostly class individuals. It corresponds to A-Box in Description Logic.

- A **Semantic Web Document** (SWD) is an online document that has an Reference ontology and may be some related annotation.

- A **Specific Semantic Web Document** (SSWD) is an online document that has an Reference ontology and may be some related annotation.
Introduction

- Semantic web has some distinguishing features that affect search process:
  - Instead of web documents, in the SW, all objects of the real world are involved in the search.
  - Information in SW is understandable by machines as well as human.
  - SW languages are more advanced than html.
  - It is possible to distribute information about a single concept in SW.
Introduction

- fundamental differences between semantic web search engines traditional search engines:
  - Using a logical framework lets more intelligent retrieval possible
  - There are more complex relations in documents
  - Specifying relationships among objects explicitly highlights the need for better visualization techniques for the results of a search.

- One important aspect of SW search is the usage of ontology and meta-data.
# Table of Content

- Introduction
- **Semantic web Search Engines**
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
A Categorization Scheme for SWSEs

- Respecting the kinds of search in SW, it is possible to categorize users to two groups.
  - Ordinary users
  - Semantic Web Application Developer

- Accordingly we can categorize SWSEs to the following two categories:
  - Engines for specific semantic web documents (SSWD, like Ontologies)
    - They search only documents that are represented in one of the languages specific to SW.
  - Engines that tries to improve search results using SW standards and languages
A Categorization Scheme for SWSEs

- **Ontology Search Engines** (Search For Ontologies)
  - ontology meta search engines
  - crawler based ontology search engines

- **Semantic Search Engines** (Search Using Ontologies)
  - Context Based Search Engines
  - Evolutionary Search Engines
  - Semantic Associations Discovery Engines
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
Ontology Search Engines

- It is not possible to use current search engines for ontologies, because:
  
  - Current techniques do not let to index and retrieve semantic tags
  - They don’t use the meaning of tags
  - Can’t display results in visual form
  - Ontologies are not separated entities and usually they have cross references which current engines don’t process
Ontology Search Engines

In general there are two approaches to handle these documents:

- Using current search engines with some modifications (Meta Searching)
- Creating a special search engines (Crawler Based Searching)
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
Ontology Meta Search Engines

- This group do retrieval by putting a system on top of a current search engine

- There are two types of this systems
  - Using Filetype feature of search engines
  - Swangling
Filetype Feature

- Google started indexing RDF documents some time in late 2003
- In the first type, there is a search engine that only searches specific file types (e.g. RSS, RDF, OWL)
- In fact we just forward the keywords of the queries with filetype feature to Google
- The main concern of such systems is on the visualization and browsing of results
OntoSearch

- A basis system with Google as its “heart”
- Abilities:
  - The ability to specify the types of file(s) to be returned (OWL, RDFS, all)
  - The ability to specify the types of entities to be matched by each keyword (concept, attribute, values, comments, all)
  - The ability to specify partial or exact matches on entities.
  - Sub-graph matching eg concept animal with concept pig within 3 links; concepts with particular attributes
Ontology Meta Search Engines

- In the second type we use traditional search engines again.
- But since semantic tags are ignored by the underlying search engine, an intermediate format for documents and user queries are used.
- A technique named Swangle is used for this purpose.
- With this technique RDF triples are translated into strings suitable for underlying search engine.
Swangling

- Swangling turns a SW triple into 7 word like terms
  - One for each non-empty subset of the three components with the missing elements replaced by the special “don’t care” URI
  - Terms generated by a hashing function (e.g., SHA1)

- Swangling an RDF document means adding in triples with swangle terms.
  - This can be indexed and retrieved via conventional search engines like Google

- Allows one to search for a SWD with a triple that claims “Ossama bin Laden is located at X”
A Swangled Triple

<rdf:RDF
   xmlns:s="http://swoogle.umbc.edu/ontologies/swangle.owl#"
></rdf>

<s:SwangledTriple>
   <rdfs:comment>Swangled text for
       [http://www.xfront.com/owl/ontologies/camera/#Camera,
       http://www.w3.org/2000/01/rdf-schema#subClassOf,
       http://www.xfront.com/owl/ontologies/camera/#PurchaseableItem]
   </rdfs:comment>
   <s:swangledText>N656WNTZ36KQ5PX6RFUGVKQ63A</s:swangledText>
   <s:swangledText>M6IMWPWIH4YQI4IMGZYBGPYKEI</s:swangledText>
   <s:swangledText>HO2H3FOPAE53AQIZ6YVPFQ2XI</s:swangledText>
   <s:swangledText>2AQEUJOYPMXWKHZTENIJS6P6M</s:swangledText>
   <s:swangledText>IIVQRXOAYRH6GGRZDFXKEEB4PY</s:swangledText>
   <s:swangledText>75Q5Z3BYAKRPLZDLFNS5KKMTOY</s:swangledText>
   <s:swangledText>2FQ2YI7SNJ7OMXOXIDEEE2WOZU</s:swangledText>
</s:SwangledTriple>
Swangler Architecture

Local KB

Semantic Web Query -> Inference Engine

Inference Engine -> Semantic Markup

Semantic Markup -> Encoder ("swangler")

Encoder ("swangler") -> Encoded Markup

Encoded Markup -> Web Search Engine

Web Search Engine -> Ranked Pages

Ranked Pages -> Extractor

Extractor -> Semantic Markup

Semantic Markup -> Filters

Filters -> Semantic Markup

Semantic Markup
What’s the point?

- We’d like to get our documents into Google
  - Swangle terms look like words to Google and other search engines.
- On the other side, this translation is done for user queries too.
  - Add rules to the web server so that, when a search spider asks for document X the document swangled(X) is returned
  - A swangle term length of 7 may be an acceptable length for a Semantic Web of $10^{10}$ triples -- collision prob for a triple $\sim 2*10^{-6}$.
  - We could also use Swanglish – hashing each triple into N of the 50K most common English words
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Conclusion
- References
Swoogle Architecture

Swoogle 2: 340K SWDs, 48M triples, 5K SWOs, 97K classes, 55K properties, 7M individuals (4/05)
Swoogle 3: 700K SWDs, 135M triples, 7.7K SWOs, (11/05)
Crawler Based Ontology Search Engines

- **Discovery**
  - Crawling of SW documents is different from html documents
  - In SW we express knowledge using URI in RDF triples. Unlike html hyperlinks, URIs in RDF may point to a non existing entity
  - Also RDF may be embedded in html documents or be stored in a separate file.
Semantic Web Crawler

- Such crawlers should have the following properties
  - Should crawl on heterogeneous web resources (owl, oil, daml, rdf, xml, html)
  - Avoid circular links
  - Completing RDF holes
  - Aggregating RDF chunks
Example of Ontology Aggregation

A distributed network of data!

Aggregated Data!
Metadata Creation

- Web document metadata
  - When/how discovered/fetched
  - Suffix of URL
  - Last modified time
  - Document size
- SSWD metadata
  - Language features
    - OWL species
    - RDF encoding
  - Statistical features
    - Defined/used terms
    - Declared/used namespaces
    - Ontology Ratio
  - Ontology Rank
- Ontology annotation
  - Label
  - Version
  - Comment
- Related Relational Metadata
  - Links to other SWDs
    - Imported SWDs
    - Referenced SWDs
    - Extended SWDs
    - Prior version
  - Links to terms
    - Classes/Properties defined/used
Digesting

- Digest
  - But the main point is that count, type and meaning of relations in SW is more complete than the current web

<table>
<thead>
<tr>
<th>Priority (Weight)</th>
<th>Relationship</th>
<th>Language Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>instantiation</td>
<td>rdf:type</td>
</tr>
<tr>
<td>2</td>
<td>subClass</td>
<td>rdfs:subclass, daml:subClass</td>
</tr>
<tr>
<td>3</td>
<td>domain/range</td>
<td>rdfs:domain, daml:range</td>
</tr>
</tbody>
</table>

Table 1: Weights of Hyperlinks
Navigating the HTML web is simple; there’s just one kind of link. The SW has more kinds of links and hence more navigation paths.
We navigate the Semantic Web via links in the physical layer of RDF documents and also via links in the “logical” layer defined by the semantics of RDF and OWL.
Rank has its privilege

- Google introduced a new approach to ranking query results using a simple “popularity” metric.
  - It was a big improvement!

- Swoogle ranks its query results also
  - When searching for an ontology, class or property, wouldn’t one want to see the most used ones first?

- Ranking SW content requires different algorithms for different kinds of SW objects
  - For SWDs, SWTs, individuals, “assertions”, molecules, etc…
Ranking SWDs

- For offline ranking it is possible to use the references idea of PageRank.
- In OntoRank values for each ontology is calculated very similar to PageRank in traditional search engines like google.
- Ranking based on “Referencing”
  - identify and rank of referrer
  - Number of citation by others
  - Distance of reference from origin to target
- Types of links:
  - Import
  - Extend
  - Instantiate
  - Prior version
  - ..
An Example

http://www.w3.org/2000/01/rdf-schema
  wPR =300  OntoRank =403

http://xmlns.com/wordnet/1.6/
  wPR =3  OntoRank =103

http://xmlns.com/foaf/1.0/
  wPR =100  OntoRank =100

http://www.cs.umbc.edu/~finin/foaf.rdf
  wPR =0.2  OntoRank =0.2
Crawler Based Ontology Search Engines

- Service
  - User interface
  - Services to application systems
Find “Time” Ontology

We can use a set of keywords to search ontology. For example, “time, before, after” are basic concepts for a “Time” ontology.

1 - 20 of total 26 results

http://www.ai.sri.com/daml/ontologies/time/Time.daml
Classes defined: 15 Properties defined: 42 Instances defined: 18
Triples: 264 Namespaces used: 5 Ontology Ratio: 0.537736
Cached: Original File N-Triples
Swoogle view: Document Properties Term Properties
Digest “Time” Ontology (document view)

[Image of a webpage showing the URI for the Time ontology: http://www.ai.sri.com/daml/ontologies/time/Time.daml]
Summary

2004

- Swoogle (Mar, 2004)
  - Automated SWD discovery
  - SWD metadata creation and search
  - Ontology rank (rational surfer model)
  - Swoogle watch
  - Web Interface

- Swoogle2 (Sep, 2004)
  - Ontology dictionary
  - Swoogle statistics
  - Web service interface (WSDL)
  - Bag of URIref IR search
  - Triple shopping cart

2005

- Swoogle3 (July 2005)
  - Better (re-)crawling strategies
  - Better navigation models
  - Index instance data
  - More metadata (ontology mapping and OWL-S services)
  - Better web service interfaces
  - IR component for string literals
Supporting Semantic Web Developers

- Finding SW content
  - Ontologies, classes, properties, molecules, triples, partial ontology mappings, authoritative copies
  - Ad hoc data collection

- Exploring how the SW is being used, e.g.
  - Computing basic statistics
  - Ranking properties used with foaf:person

- And misused
  - Finding common typos
Applications and use cases

- Supporting Semantic Web developers, e.g.,
  - Ontology designers
  - Vocabulary discovery
  - Who’s using my ontologies or data?
  - Etc.

- Searching specialized collections, e.g.,
  - Proofs in Inference Web
  - Text Meaning Representations of news stories in SemNews

- Supporting SW tools, e.g.,
  - Discovering mappings between ontologies
# Table of Content

- **Introduction**
- **Semantic web Search Engines**
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- **Discussion and Evaluation**
- **References**
There are some restrictions for current search engines

One interesting example: "Matrix"

Another example is java

Semantic web is introduced to overcome this problem.

The most important tool in semantic web for improving search results is context concept and its correspondence with Ontologies. This type of search engines uses such ontological definitions
Two Levels of the Semantic Web

- **Deep Semantic Web:**
  - Intelligent agents performing inference
  - Semantic Web as distributed AI
  - Small problem … the AI problem is not yet solved

- **Shallow Semantic Web:** using SW/Knowledge Representation techniques for
  - Data integration
  - Search
  - Is starting to see traction in industry
Problems with current search engines

- Current search engines = keywords:
  - high recall, low precision
  - sensitive to vocabulary
  - insensitive to implicit content
Semantic Search Engines

- It is possible to categorize this type of search engines to three groups.
  - Context Based Search Engines
    - They are the largest one, aim is to add semantic operations for better results.
  - Evolutionary Search Engines
    - Use facilities of semantic web to accumulate information on a topic we are researching on.
  - Semantic Association Discovery Engines
    - They try to find semantic relations between two or more terms.
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
    - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
Context Based Search Engines
Context Based Search Engines

1) Crawling the semantic web:

- There is not much difference between these crawlers and ordinary web crawlers.
- Many of the implemented systems use an existing web crawler as an underlying system.
- It's better to develop a crawler that understands special semantic tags.
- One of the important features of these crawlers should be the exploration of ontologies that are referred from existing web pages.
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
Annotation Methods

- Annotation is perquisite of Search in semantic web.
- There are different approaches which spawn in a broad spectrum from complete manual to full automatic methods.
- Selection of an appropriate method depends on the domain of interest
- In general meta-data generation for structured data is simpler
Annotation Methods

Annotations can be categorized based on following aspects:

- **Type of meta-data**
  - Structural: non contextual information about content is expressed (e.g. language and format)
  - Semantic: The main concern is on the detailed content of information and usually is stored as RDF triples
Annotation Methods

- Generation approach
  - A simple approach is to generate meta-data without considering the overall theme of the page. (Without Ontology)
  - Better approach is to use an ontology in the generation process.
    - Using a previously specified ontology for that type, generate meta-data that instantiates concepts and relations of ontology for that page
    - The main advantage of this method is the usage of contextual information.
Annotation Methods

- **Source of generation**
  - The ordinary source of meta-data generation is a page itself
  - Sometimes it is beneficial to use other complementary sources, like using network available resources for accumulating more information for a page
    - For example for a movie it might be possible to use IMDB to extract additional information like director, genre, etc.
Context Based Search Engines

- Knowledge Parser is a kind of complete system using important techniques
3) Indexing:

Most of the engines does not provide any special functionality regarding indexing.

OWLIR uses Swangling explained earlier.

Also in DOSE possibility of dividing documents to smaller parts is used to improve indexing performance.

Also in one of p2p architecture Semantic Searching, for each of concepts in the reference ontology there exist an agent that maintains information corresponding to it.
QuizRDF Introduces Ontological Indexing in which indexing is done based on a reference ontology.
4) Accepting user’s requests:

- There are two different approaches:
  - term-based
  - form-based.
- In term-based approach it is tried to find the search context from entered keywords.
- In the form-based approach user interface is generated according to the ontology selected by user.
Context Based Search Engines

5) Generating meta data for user requests:
   - This operation is very similar to generating metadata for documents.
   - For example in DOSE the same Semantic Mapper is used for generating metadata both for documents and user requests.
   - Often Wordnet is used to expand user requests.
   - For example for termed entered by a user, using Wordnet, synonyms can be extracted and used to expand the query.
Context Based Search Engines

6) Retrieval and ranking model:

- Usually an ordinary VSM model is used then based on RDF graph matching results are pruned.
- From the equivalence of RDF graphs and Conceptual Graphs (CG), already existing operations on CGs is used to match user request and documents.
- Semantic Distance concept is often used to estimate similarity of concepts in a matching process.
- It is also possible to use graph similarity for ranking results.
- Fuzzy approach is used for this purpose too
Context Based Search Engines

7) Display of results:
- A major different of semantic search engines and ordinary ones is the display of results.
- One of the primary tasks is to filter the results (for example for eliminating repetitions).
- In QuizRDF in addition to normal display of results, a number of classes is displayed and when a user selects one, only those results having instances of those class is shown.
- display is a kind of hierarchy in which top concepts of ontology is shown and by selecting one its children detail of it according to the ontology is displayed.
QuizRDF

- combined text- and ontology-based search engine
- low-threshold, high-ceiling
QuizRDF

Results: matching: energy

1. IT in Energy
   - access, agent, Akkermans, application, approach, area, assessment, auction, bandwidth, business, capability, change, channel, communication, community, company, competition, contact, control, cost, customer, December, December,
17. Power Line Telecommunications Report
line, navigation, page, page navigation panel, panel, power, Publication, report, Resource
hasId: 26
hasAuthor: Healey, David
hasAuthor: Akkermans, Hans
hasAuthor: Ottersten, Hans
hasDate: 6/15/96 0:00:00
hasProject: SES7/Power line communication
hasTitle: Transmission of Data over the Electricity Power Lines

18. Power Line Telecommunications Report
access, agent, application, area, business, capability, communication, control, cost, customer, development, distribution, efficiency, electricity, energy, energy system, equipment, example, expansion, form, grid, home, ICT, industry, information, infrastructure, line, local, management, may, office, opportunity, order, page, participation, PLT, power, report, Resource, sector, service, society, system, technology, utility, value

19. Power Line Telecommunications Report
approach, area, centre, change, communication, condition, control, customer, deregulation, development, efficiency, electricity, energy, example, grid, ICT, industry, information, information system, infrastructure, line, management, market, page, part, pilot, PLT, power, power line, product, Resource, sector, service, study, system, time, today, transactions, utility, way

20. Impact of Deregulation - DSM in Industries
commodity, distribution, efficiency, energy, energy utility, grid, hour, line, load, load management, management, power, power line, Resource, utility
Main Search Grid Session Impact of Deregulation - DSM in Industries Ulfka Bergström, Sydkraft Konsult AB and the Linköping University, Sweden Summary The composition of the deregulated (energy x51)
QuizRDF

Results: restricted to Publication, matching: energy

1. IT in Energy
   Publication
   IT in Energy - A SITI National Research Program proposal - 1. Summary: - The global deregulation of the telecommunication markets, the opening of the energy sector, the decreasing costs of (energy x 2)
QuizRDF

Results: restricted to Publication, matching: energy

1. IT in Energy
   Publication
   IT in Energy - A SITI National Research Program proposal - 1. Summary: - The global deregulation of the telecommunication markets, the opening of the energy sector, the decreasing costs of (energy x 82)
   hasTitle: 200
   hasAuthors: multiple
Searching The Semantic Web

Semantic Web Research Laboratory
Sharif University of Technology

QuizRDF

1. Structured Engineering Process for Agent Communication Modelling
   Publication
   Structured Engineering Process for Agent Communication Modelling Hans Akkermans & Rune Gustavsson & Fredrik Ygge
   AKMC and Free University Amsterdam Volume.net (energy: 18, Akkermans: 21)
   hasTitle: Structured Engineering Process for Agent Communication Modelling
   hasPage: https://143.227.135.22/evaluator/knowledgebase/publications/conference-journals/ismik96/ismik96-2.html

2. HOMEBOOTS: Intelligent Decentralized Services for Energy Management
   Publication
   HOMEBOOTS: Intelligent Decentralized Services for Energy Management Hans Akkermans Fredrik Ygge Rune Gustavsson AKMC
   Knowledge Management and University of Twente Information Systems Department (energy: 32, Akkermans: 6)
   hasTitle: HOMEBOOTS: Intelligent Decentralized Services for Energy Management
   hasPage: https://143.227.135.22/evaluator/knowledgebase/publications/conference-journals/ismik96/ismik96-2.html

3. Power Line Telecommunications Report
   Publication
   hasTitle: Transmission of Data over the Electricity Power Lines
   hasPage: https://143.227.135.22/evaluator/knowledgebase/publications/conference-journals/ismik96/ismik96-2.html
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
Evolutionary Search Engines

- The advanced type of search is something like research.
- Here we aim at gathering some information about a specific topic.
- It can be something like search by Teoma search engine.
- For example, if we give the name of a singer to the search engine, it should be able to find some related data to this singer like biography, posters, albums, and so on.
Evolutionary Search Engines

- These engines usually use one of the commercial search engines as their base component for searching and they augment returned results by these base engines.
- This augmented information is gathered from some data-insensitive web resources.
Evolutionary Search Engines Architecture

- Web Pages
- Informational Web Sites
- Annotator
- Context Detector
- Annotation Reliever
- Annotation Repository
- Additional Information for Query
- Query
- User
- Traditional Search Engine
- Main Concepts
- Raw Result
- Displayed Results

Searching The Semantic Web
Semantic Web Research Laboratory
Sharif University of Technology
Evolutionary Search Engines

- It has some similarities with previous category’s architecture
- Here we crawl and generate annotation just for some well know informational web pages i.e. CDNow, Amazon, IMDB
- After this phase we collect annotations in a repository.
Whenever a sample user posed a query to processes must be performed:

- first, we should give this query to a usual search engine (usually Google) to obtaining raw results.
- Second, system will attempt to detect the context and its corresponding ontology for the user’s request in order to extract some key concepts.
- Later we use these concepts to fetch some information from our metadata repository.
- The last step in this architecture is combining and displaying results.
Evolutionary Search Engines

- Main problems and challenge in these types of engines are:
  - Concept extraction from user’s request
  - Selecting proper annotation to show and their order
Evolutionary Search Engines

- Concept extraction from user’s request
  - there are some problems that lead to misunderstanding of input query by system;
    - Inherent ambiguity in query specified by user
    - Complex terms that must be decomposed to understand.
Evolutionary Search Engines

- Selecting proper annotation to show and their order:
  - often we find a huge number of potential metadata related to the initial request and we should choose those ones that are more useful for user.
  - A simple approach is using other concepts around our core concept (which we extracted it before) in base ontology
  - if we have more than one core concept we must focus on those concepts that are on the path between these concepts.
Displaying the Results

- Results are displayed using a set of templates
- Each class of object has an associated set of templates
- The templates specify the class and the properties and a HTML template
- A template is identified for each node in the ordered list and the HTML is generated
- The HTML is included in the results page
W3C Search

- W3C Semantic Search has five different data sources: People, Activities, Working Groups, Documents, and News
- Both the ABS and W3C Semantic Search have a basic ontology about people, places, events, organizations, vocabulary terms, etc.
- The plan is to augment a traditional search with data from the Semantic Web
A segment of the Semantic Web pertaining to Eric Miller
Sample Applications-W3C Search
Activity Based Search

- ABS contains data from many sites, such as AllMusic, Ebay, Amazon, AOL Shopping, TicketMaster, Weather.com and Mapquest
- There are millions of triples in the ABS Semantic Web
- TAP knowledge base has a broad range of domains including people, places, organizations, and products
- Resources have a rdf:type and rdfs:label
Search Augmenter

Yo-Yo Ma is a Musician

SS KB

Sample Applications-ABS

Searching The Semantic Web

Sharif University of Technology
Sample Applications-ABS

Activity Based Search

The results on the left hand side of the page are from Google. The boxes on the right hand side of the page, if any, have been drawn in real time from the Semantic Web. These boxes [if any] were chosen based on the denotation of the search term in the TAP KB. Click on the term to be taken to the appropriate entry in the TAP KB.

Michael Jordan Official Website

... Jordan fights a career with addiction in Philly. Michael Jordan's final shot was a free throw, and it's his final appearance in an NBA uniform. It was good. ... Description: Learn about basketball, Mike's career, training tips and Mike's favorite charities.
Category: Sports and Teams > Entertainment > Celebrities
jordan.spordline.com - 31k - Cached - Similar pages

Michael Jordan's Home Page

Michael Jordan Professor University of California at Berkeley, Computer Science 321 Soda Hall #1776 Berkeley, CA 94720-1776 Phone ... Description: Graphical models, variational methods, machine learning, reasoning under uncertainty.
Category: Computers > Artificial Intelligence > Neural Networks > People
www.cs.berkeley.edu/~jordan - 5k - Cached - Similar pages
www.nba.com/playerfiles/michael_jordan.html - 2k - Cached - Similar pages

NBA.com Michael Jordan Player Info
... Michael Jordan 23 ... www.nba.com/playerinfo/michael_jordan/snap.html - 99k - Cached - Similar pages
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
Semantic Association Discovery SEs

- Usually one of the user’s interests is finding semantic relations between two input terms
- The focus is to expand search to include relationship search in addition to document search
- To be able to ask a query like “how are entities x and y related”
  - eg in case of investigative domain, we should be able to ask a query like how are two passengers X and Y related
Semantic Association Discovery SEs

- Old search engines handled these requests using learning and statistical methods.
- Semantic web standards and languages have provided more effective and precise methods.
- There are different types of semantic associations.
- Usually we talk about just two terms because the average length for users’ queries is 2.3 terms.
Knowledge-based Associations

Islamic Jihad vows 'revenge'

It is not yet known who was responsible for the bombing.

However, in the past, the Islamic Jihad movement was led by "revenge" against the US for its involvement in the extradition to Egypt of several militants living in eastern Europe.

Here are excerpts from a report by an Arabic newspaper Al-Hayat of August 6:

Cairo: The "Islamic Jihad Group," led by Dr Ayman al-Zawahiri, is known to have taken revenge against the US for its involvement in the extradition to Egypt of several militants living in eastern Europe.

There are a number of Egyptian citizens among them, including a high-ranking figure in the organization, Ahmed Ibrahim al-Nagger, who was sentenced to death by the High Court in Cairo in connection with the Khan al-Khalili case.

Ahmed Ibrahim al-Nagger, one of 12 handed over to Egypt from Albania last summer, was the red clothing of a condemned man because he had been sentenced to death in an earlier trial for plotting an attack on Cairo's Khan al-Khalili market.
Examples in 9-11 context

- What are relationships between Khalid Al-Midhar and Majed Moqed?
  - **Connections**
    - Bought tickets using same frequent flier number
  - **Similarities**
    - Both purchased tickets originating from Washington DC paid by cash and picked up their tickets at the Baltimore-Washington Int'l Airport
    - Both have seats in Row 12

- “What relationships exist (if any) between Osama bin Laden and the 9-11 attackers”
Semantic Associations From Graph

Artist
creates
Artifact
exhibited
Museum

Sculptor

Sculpture

Painter

Painting

Ext. Resource

String
last_modified
Date

String
file_size

String

title

String
mime-type

String
subClassOf(isA)

String
subPropertyOf

String
typeOf(instance)

“Pablo”
name

“Picasso”
name

“Rembrandt”
name

“Rodin”
name

“Rodin Museum”
title

“Reina Sofia Museum”
title

2000-6-09
last_modified

creates

paints

exhibited

paints

exhibited

paints

exhibited

Closed-World Assumption

Searching The Semantic Web

Sharif University of Technology
Semantic Associations From Graph

- &r1 and &r3 have an association because &r1 paints a painting (&r2) which is exhibited at the museum (&r3).

- &r4 and &r6 are semantically associated because they both have created artifacts (&r5, and &r7) which are exhibited at the same museum (&r8).

- &r1 and &r6 are associated because of a similarity in their relationships. For example, they both have creations (&r2, and &r7) that are exhibited by a Museum (&r3, &r8).
ρ - Association

- Two entities $e_1$ and $e_n$ are semantically connected if there exists a sequence $e_1, P_1, e_2, P_2, e_3, \ldots e_{n-1}, P_{n-1}, e_n$ in an RDF graph where $e_i$, $1 \leq i \leq n$, are entities and $P_j$, $1 \leq j < n$, are properties.
Two entities are semantically similar if both have $\geq 1$ similar paths starting from the initial entities, such that for each segment of the path:

- Property $P_i$ is either the same or subproperty of the corresponding property in the other path
- Entity $E_i$ belongs to the same class, classes that are siblings, or a class that is a subclass of the corresponding class in the other path
Semantic Association

- \( \rho \) - Query
  - A \( \rho \) - Query, expressed as \( \rho(x, y) \), where \( x \) and \( y \) are entities, results in the set of all semantic paths that connect \( x \) and \( y \)

- \( \sigma \) - Query
  - A \( \sigma \) - Query, expressed as \( \sigma(x, y) \), where \( x \) and \( y \) are entities, results in the set of all pairs of semantically similar paths originating at \( x \) and \( y \)
Discovery Techniques

- For finding semantic association between input terms some techniques have been proposed:
  - Bayesian networks:
    - graph and parameters
  - Spread Activation Technique:
    - we can expand an initial set of instances to contain most relative instances to them.
    - The initial set is populated by extracting important terms from user’s query, then with respect to the metadata repository corresponding instances is retrieved and after expanding this instance an instances graph is produced
Ranking Semantic Associations

- After discovery phase often we have numerous semantic association, therefore a ranking policy must be used.
- i.e. for Terrorism test bed with > 6,000 entities and > 11,000 explicit relations.
- The following semantic association query $\rho(\text{"Nasir Ali"}, \text{"AlQeada"})$, results in 2,234 associations.
- The results must be presented to a user in a relevant fashion…thus the need for ranking.
Ranking Semantic Associations

- Semantic metrics
  1. Context
  2. Subsumption
  3. Trust

- Statistical metrics
  1. Rarity
  2. Popularity
  3. Association Length
Context

- Context => Relevance; Reduction in computation space
- Context captures the users’ interest to provide the user with the relevant knowledge within numerous relationships between the entities
- By defining regions (or sub-graphs) of the ontology we are capturing the areas of interest of the user
Context Weight

- Consider user’s domain of interest (user-weighted regions)

- Issues
  - Paths can pass through numerous regions of interest
  - Large and/or small portions of paths can pass through these regions

- Paths outside context regions rank lower or are discarded
Context Weight - Example

Region1: Financial Domain, weight=0.50
Region2: Terrorist Domain, weight=0.75
Subsumption Weight

- Specialized instances are considered more relevant
- More “specific” relations convey more meaning
Path Length Weight

- Interest in the most direct paths (i.e., the shortest path)
  - May infer a stronger relationship between two entities

- Interest in hidden, indirect, or discrete paths (i.e., longer paths)
  - Terrorist cells are often hidden
  - Money laundering involves deliberate innocuous looking transactions
Path Length - Example

- Short Paths Favored
- Ranked Higher (0.889)
- Ranked Lower (0.01)

- Long Paths Favored
- Ranked Higher (1.0)
- Ranked Lower (0.1111)
Trust Weight

- Relationships (properties) originate from differently trusted sources
- Trust values need to be assigned to relationships depending on the source
- e.g., Reuters could be more trusted than some of the other news sources
- Current approach penalizes low trusted relationships (may overweight lowest trust in a relationship)
Ranking Criterion

- Overall *Path Weight* of a semantic association is a linear function

\[
\text{Ranking Score} = k_1 \times \text{Subsumption} + k_2 \times \text{Length} + k_3 \times \text{Context} + k_4 \times \text{Trust}
\]

where \( k_i \) add up to 1.0

- Allows fine-tuning of the ranking criteria
Sample Application - SemDis
# Table of Content

- **Introduction**
- **Semantic web Search Engines**
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- **Discussion and Evaluation**
- **References**
Discussion and Evaluation

- Unfortunately most of these search engines has implemented through the research projects and therefore they are not available for testing and evaluating.
- In the other hand because of their differences with traditional search engine it’s not possible to compare them using unique evaluation framework.
- Here we mention just some points and hints for comparing and evaluating these search engines based on our categorization scheme.
Ontology Meta Search Engines

- The main goal is finding SWDs specially ontologies
- We use traditional search engines for this purpose
- There are two approaches in using usual search engines:
  - search only by the name of files and use some options like filtetype (rdf,owl,rss,..)
  - search by labels by converting both documents and queries to intermediate format that is not ignorable for ordinary search engines.
Ontology Meta Search Engines

- Having a good display module for browsing and navigating the founded ontologies is critical point
- Examples:
  - Swangler[2]
  - OntoSearch[8]
Crawler Based Ontology Search Engines

- Here we use a specific crawler to find SWDs on the web, index them and extract some metadata about them.
- By using the engines we can search by special class or property and even for sample data (ABox).
- Graph structure of the SWDs on the web can be explored by use of these search engines.
Crawler Based Ontology Search Engines

- Also here visualizing the results is important.
- Examples:
  - Swoogle[2,3,4]
  - Ontokhoj[27]
Ontology Search Engines

- In contrast to usefulness of meta-search engines for regular pages in traditional web, it seems that they are not so good for ontologies.
- In fact we can not collect the all ontologies in the web just but using filetype command within commercial search engines.
- In addition swangling operation has a huge amount of overhead.
Ontology Search Engines

- It’s much better to use crawler-based ontology search engines (2nd category) rather than ontology meta-search engines (1st category).
- In order to evaluating performance of this kind of search engines there is no standard test collection.
- We can simply Evaluate them by searching for ontologies using:
  - label of ontologies
  - classes
  - and properties
Ontology Search Engines

- Benchmarking and developing an ontology test collection for these search engines is an open problem.
- Ontology Repositories can be useful in this area.
- DAML Ontology library:
  - 282 ontologies
  - total no. of classes 67987
  - total no. properties 11149
  - total no. of instances 43646
Context Based Search Engines

- Purpose of these engines is enhancing performance of traditional search engines
- These engine are the most practical ones
- They are most popular search engines in the semantic web
- The main strangeness of these engines is their simplicity
- In fact they tried to be as simple as textbox search engines (like google)
Context Based Search Engines

- Gaining better results is possible through understanding the context of documents and queries (using of ontologies)
- One of the important part of this type is annotator which responsible for generating metadata for crawled pages.
- We need to generate some metadata for user’s query too
- After traditional retrieval we combine matching RDF graphs to obtain better quality of results.
Context Based Search Engines

- The biggest problem of these search engines is that they are limited to the special contexts.
- It’s very better if we can develop a multi-context semantic search engine.
Context Based Search Engines

- Fortunately we can apply standard measures (i.e. Precision and Recall) and test collections (i.e. TREC tracks) of traditional information retrieval to evaluate this kind of semantic web search engines.

- It should be noticed that if we can prepare ontology for test documents, the results will show much improvements.

- Examples
  - OWLIR[2], QuizRDF[6], InWiss[7],
  - Corese[9], Infofox[12], SHOE[15],
  - DOSE[18], SERSE[22], ALVIS[17],
  - OntoWeb[23], Score[25], [20],[21], [24]
Evolutionary Search Engines

- This type of search engines aim at information gathering for user’s request
- We can suppose these engines as the semantic type of HITS-based search engines (i.e. Teoma) which exploit hub and authority pages for user’s request
- They usually use an ordinary search engine and display augmented information near the original results
- They use external metadata
Evolutionary Search Engines

- This category of search engines is usually specific for special application domains.
- In a large-scale mode like (i.e. in whole web) they will be very similar to a multi context search engines.
- Examples:
  - W3C Semantic Search[5]
  - ABS[5]
Semantic Association Discovery SEs

- The goal is finding various semantic relations between input terms (usually two) and then rank the results based on semantic distances metrics.
- They are more adaptable with knowledge Bases
- Compared to other categories, the semantic association discovery engines are related to higher layers of semantic web cake (logic and proof).
- Result of these engines is very depending on their ontology repository
Semantic Association Discovery SEs

- An upper ontology like WordNet or OpenCyc can be used for evaluating this kind of search engines.
- After selecting two concepts randomly, the correctness and speed of discovering paths between them are two useful measures for performance evaluation.
- Examples:
  - SemDis[10,14]
  - [13]
  - [16]
Table of Content

- Introduction
- Semantic web Search Engines
  - Ontology Search Engines
    - Meta Ontology Search Engines
    - Crawler Based Ontology Search Engines
  - Semantic Search Engines
    - Context Based Search Engines
      - Semantic Annotation
    - Evolutionary Search Engines
    - Semantic Association Discovery Engines
- Discussion and Evaluation
- References
References


References

References


References


Semantic Web Research Laboratory

Searching The Semantic Web

Sharif University of Technology

128