From Web Search to Semantic Search

XIANG ZHANG
x.zhang@seu.edu.cn
Content

- Web and Semantic Web
- Two faces of Semantic Web
- Web Search
- Semantic Search – State of Art
- Falcons
- Searching Semantic Associations
"A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."
“Our wearable data collection system lets users collect their experiences into a continually growing and adapting multimedia diary. The system—called inSense—uses the patterns in sensor readings from a camera, microphone, and accelerometers to classify the user’s activities and automatically collect multimedia clips when the user is in an "interesting" situation.”
• Memex is internally complex in mechanics
  o dry photography, microphotography...
  o photocells, thermionic tubes, cathode ray tubes...
• Memex is externally simple for human usage
  o associative indexing
    ▪ tying two items together and name it
    ▪ linking enormous items to form a trail and name it
Building Internet in the Cold War
Building Internet in the Cold War

- The idea of package switching (early 1960s)
- APARNet project launched (1960s) – US Gov
  - robust / fault-tolerant / distributed
- APARNet with 4 nodes (Univ. Nodes) (1969)
- The idea of OSI (Open System Interconnection) (1970s)
- NSFNet (National Science Foundation) (1986)
- APARNet closed (1990)
- >1 million hosts in Internet (1992)
Building Internet in the Cold War

- why Internet becomes a successful creation
  - divide-and-conquer
    - package switching
  - abstraction
    - OSI
  - Scalability / reliability > speed / efficiency
    - TCP / IP
Ted Nelson and His Hypertext

I don’t like Tim Berners-Lee and his WWW. Too Simple!!!
Ted Nelson and His Hypertext

- his first job is a photographer and film editor
- “Hypertext” is coined by him in 1963
  - based on files
  - data was stored once
  - no deletion
  - information was accessible by a link from anywhere
Ted Nelson and His Hypertext

- other achievement of Ted
  - Project Xanadu (1960)
  - found HTTP Company
  - produced the first bag for the first laptop
  - now producing 47% bags for laptop on the world
Doug Engelbart

- other achievement of Doug (Tuning Award)
  - mouse (commercially implemented in LISA 1983)
  - Windows
  - shared-screen teleconferencing
  - hypermedia
  - groupware
Ted and Doug

- they are both really avant-garde
- every technology was in place in 1970s
- but it took the **PC revolution** and widespread **internet** to inspire the **WWW**
- the process lasted for almost **20** years!
Tim and His Web Dream

Tweets

Tim Berners-Lee @timberners_lee
This is for everyone #london2012 #oneweb #openingceremony
@webfoundation @w3c
Collapse  Reply  Retweet  Favorite

8,622 RETWEETS  1,192 FAVORITES

10:08 PM - 27 Jul 12 via Twitter for iPhone - Details

Pocket lint

THIS IS FOR EVERYONE!
Why we use HTML for hypertext representation?

- HTML vs. XML

```html
<html>
  <title>I am Zhang Xiang</title>
  <body>
    <p>This is a paragraph</p>
    <p>This is another paragraph</p>
  </body>
</html>
```

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<note>
  <to>George</to>
  <from>John</from>
  <heading>Reminder</heading>
  <body>Don't forget the meeting!</body>
</note>
```
Tim and His Web Dream

- HTTP merits
  - quite simple (get or post)
  - flexible (content-type)
  - weak coupling (no connection / no state)
http://www.ted.com/talks/tim_berners_lee_a_magna_carta_for_the_web
Architecture of Semantic Web

- Trust
- Digital Signature
- Proof
- Logic
- Data
- RDF + rdfschema
- XML + NS + xmschema
- Unicode
- URI

Self-desc. doc.

Ontology vocabulary
Why Semantic Web?

- **Web Intelligence**
  - Machine-readable
  - Reasoning

- **Web of Data**
  - Identification and Interlinks
  - Programmable Web
Ontology and RDF
Instantiation of Ontology
XML Representation of Ontology

```
<owl:Class rdf:about="http://xmlns.com/foaf/0.1/Person">
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
</owl:Class>

<owl:Class rdf:about="http://xmlns.com/univ/0.1/University">
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
</owl:Class>

<owl:Class rdf:about="http://xmlns.com/univ/0.1/Professor">
  <rdfs:subClassOf rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
</owl:Class>

<owl:ObjectProperty rdf:about="http://xmlns.com/univ/0.1/employedBy">
  <rdfs:domain rdf:resource="http://xmlns.com/univ/0.1/Professor"/>
  <rdfs:range rdf:resource="http://xmlns.com/univ/0.1/University"/>
</owl:ObjectProperty>
```
The Logic Face of Semantic Web

Diagram showing the relationships between key concepts such as XML Tree, RDF Graph, logic Formula, and URI.
Description Logic and Reasoning

\[ \text{Actor} \sqsubseteq \text{Artist} \]

\[ \text{hasChild} \circ \text{hasChild} \sqsubseteq \text{hasSibling} \]

\[ \text{married} \sqsubseteq \text{loves} \]

\[ \text{married(angelina, brad)} \]

\[ \text{Polygamist} \sqsubseteq \geq 2.\text{Married} \]

\[ \exists \text{Married.\{brad\}} \sqsubseteq \{\text{angelina}\} \]
The Data Face of Semantic Web
Applications of Semantic Web

- Biomedicine
- Transportation Engineering
- Homeland Security
- Software Design
- Travel Planning
- Job Finding
- Online Dating...
http://www.ted.com/talks/tim_berners_lee_on_the_next_web
http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide
Semantic Web in a Nutshell

• Similar but different with
  o Relational Database
  o Object-oriented System

• Shifted from logic to data
  o To keep it simple
  o To grow into large scale
Web Search

- 1993: Web Crawler
- 1994: Yahoo
- 1994: Lycos
- 1995: Altavista
- 1998: Google
Searching Everything

- **Google now**
  - 10 billion pages indexed in 2011 (rumor: 100 billion static)
  - average response time: 0.25s
  - 2 billion Kilowatt-hours energy in 2010
Searching Everything

- Why Google goes big?
  - when IT meets Mathematics
  - scalability, scalability and scalability
Web Search

- Challenge of Web Search
  - Distributed Data
  - Volatile Data – changed or dead data
  - Large Volume
  - Unstructured and Redundant Data
  - Data Quality
  - Data Heterogeneous
Semantic Search

• Different with Web Search
  o RDF Graphs vs. Bag of Words
  o Ontology, RDF document, Entity(object) vs. Pages
  o Ranking
  o Summarization / Snippet / Recommendation

• Similar with Web Search
  o Keyword-based
What can we get when searching “tim”

Tim.foaf.xml

Tim

Mary

hasWife

knows

Southeast University

worksAt

located

Nanjing

worksAt

Jack
More examples: Who is Bob Marley? From Falcons

Bob_Marley is a Person, Singer
- name: Marley, Bob - From dbpedia.org »
- givenname: Bob - From dbpedia.org »
- label: Bob Marley - From dbpedia.org »
http://dbpedia.org/resource/Bob_Marley - Described in 43 documents
More examples: Who is Bob Marley? From Falcons

DBpedia Properties

- about property: the comedian (en)
- about property: Bob Marley (comedian) (en)
- about property: the reggae musician (en)
- abstract: 巴布·馬利（Bob Marley，1945年2月6日 - 1981年5月11日），本名羅伯特·內斯塔·馬利（Robert Nesta Marley），是一位牙買加的歌手、吉他手、作曲家和社會運動者。他是最為人知的雷鬼歌手。他多數的作品都是經過貧困與成長的掙扎。巴布·馬利也以他透過音樂傳播信仰的方式而聞名。他是麗塔·安德森·馬利的丈夫。麗塔·安德森·馬利是巴布·馬利的伴奏組團Thees的成員，經常與他一起登台表演。" (zh)
More examples: Who is Bob Marley? From Google and Baidu

Bob Marley: Home: The Official Site - [ 翻译此页 ]
A tribute to the legendary Bob Marley exploring his life, music, and philosophy. Includes unseen photographs, essays, sound, video, and merchandise.
www.bobmarley.com - 网页快照 - 类似结果

Bob Marley - Wikipedia, the free encyclopedia - [ 翻译此页 ]
Robert Nesta "Bob" Marley (February 6, 1945 – May 11, 1981) was a Jamaican singer, songwriter, and musician. He was the lead singer, songwriter and guitarist of Track Martyn, and was the lead singer of the Wailers.
en.wikipedia.org/wiki/Bob_Marley - 网页快照 - 类似结果

雷鬼乐教父 BOB MARLEY

2007年1月27日 ... Bob Marley, 一个音乐界不容忽视的名字。他是第三世界的流行巨星, 他的音乐包含了宽容、博爱及信仰，他短暂的一生赢得了全世界的认可。 www.9sky.com/topic/index/20070127 - 网页快照 - 类似结果

雷鬼乐教父 BOB MARLEY

Marley与他乐队在雷鬼乐的形态里给予牙买加人民自尊、自信和和平。1963年他与其它乐手改名为“The Wailers”乐队。他说任何为了正义而呐喊的人都可以理解。仅36岁的 Bob Marley离开了深爱他的人们他对理想的执着, 对流行音乐、对人民群众的热爱。他把一生的时间献给了雷鬼乐。
Bob Marley:
No Woman, No Cry

Bob Marley:
Three Little Birds
More examples: Who is the tallest player in NBA? From Falcons

- **Center_%28basketball%29**
  - abstract: ...for a team. The centers are also generally the **players** who are chosen to take jump balls. Among curr...
  - has subject: Category:Centers_%28basketball%29
  - wikipage zh: [http://zh.wikipedia.org/wiki%E4%B8%AD%E9%83%B2_%28.E7%B1%83.E7%9D%83%29](http://zh.wikipedia.org/wiki%E4%B8%AD%E9%83%B2_%28.E7%B1%83.E7%9D%83%29)

- **Gheorghe Mureșan**
  - abstract: ...a retired Romanian professional basketball **player**. At 7'7" (2.31 m), he is arguably the **tallest**...
  - has subject: Category:Living_people

- **Sweedo_Halbrook**
  - abstract: ...533, April 5, 1888) was an American former **NBA player** who played from 1960-1962. He played in colleg...
  - has subject: Category:1988_deaths

- **Manute_Bol**
  - abstract: ...October 15, 1962) is a Sudanese-born basketball **player** and activist. Until the debut of Gheorghe Mureș...
  - has subject: Category:Living_people
  - wiki page uses template: Template infobox **nba player**

[http://dbpedia.org/resource/Center_%28basketball%29](http://dbpedia.org/resource/Center_%28basketball%29)
[http://dbpedia.org/resource/Gheorghe_Mure%c5%9Fan](http://dbpedia.org/resource/Gheorghe_Mure%c5%9Fan)
[http://dbpedia.org/resource/Sweedo_Halbrook](http://dbpedia.org/resource/Sweedo_Halbrook)
[http://dbpedia.org/resource/Manute_Bol](http://dbpedia.org/resource/Manute_Bol)
More examples: Who is the tallest player in NBA? From Google and Baidu
More examples: Who is chris@bizer.de? He knows who? From Falcons

Falcon's Object Concept Document
chris@bizer.de
Separate keywords with a space, and put a phrase in double quotes.

Specify a type:

Ontology Person

Objects 1-4 of 4 for your search chris@bizer.de (0.01 seconds)

! _Chris_Bizer is a Person, Thing
  - givename: Chris
  - label: Chris Bizer
  - personal mailbox: mailto chris@bizer.de
    http://social.semantic-web.at/wiki/index.php/Chris_Bizer - Described in 65 docs

! chris is a Person
  - name: Chris Bizer
  - nick: Chris Bizer
  - personal mailbox: mailto chris@bizer.de
    http://www.bizer.de/~chris - Described in 24 docs

Fresnel Lens and Format Core Vocabulary is a Ontology
  - Title: Fresnel Lens and Format Core Vocabulary
  - description: OWL Full vocabulary for defining lenses and formats on RDF models
  - Creator: Chris Bizer <chris@bizer.de>
    http://www.w3.org/2004/10/fresnel - Described in 2 docs
More examples: Who knows Chris Bizer? From Falcons

![Diagram](image)

- **Denny Vrandečić** is a *Person, Thing*
  - *givenname: Denny*
  - *label: Denny Vrandečić*
  - *knows: Chris Bizer*

- **Soeren Auer** is a *Person, Thing*
  - *givenname: Sören*
  - *label: Soeren Auer*
  - *knows: Chris Bizer*
  - [From social.semantic-web.at »](http://social.semantic-web.at/wiki/index.php/Soeren_Auer)

- **Jens Lehmann** is a *Person, Thing*
  - *givenname: Jens*
  - *label: Jens Lehmann*
  - *knows: Chris Bizer*

- **Michael Hausenblas** is a *Person, Ontology, Subject*
  - *name: Michael Hausenblas*
  - *label: Michael Hausenblas*
  - *knows: Chris Bizer*
  - [From semanticweb.org »](http://semanticweb.org/id/Michael_Hausenblas)

- **Tom Heath** is a *Person, Ontology, Subject*
  - *name: Tom Heath*
  - *label: Tom Heath*
  - *knows: Chris Bizer*
  - [From semanticweb.org »](http://semanticweb.org/id/Tom_Heath)
Semantic Search

- 2003: TAP (Guha, McCool, & Miller, 2003)
  - Earliest semantic web search engine
  - Keyword query
  - Return an object and its surrounding subgraph using label information
  - Selection is based on popularity, user profile, and search context
2005: Swoogle (Ding et al., 2005)

- One of the most popular semantic web search engine
- Class/property search and ontology search
- Pagerank-like ranking algorithm
- Provides statistical metadata of results
Swoogle
semantic web search

ontology document term across ontologies

Swoogle Search

Searching over 10,000 ontologies
## Ontology Search of Swoogle

<table>
<thead>
<tr>
<th>ontology</th>
<th>document</th>
<th>term</th>
<th>across ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### List ontologies matching ontology search

**http://swrc.ontoware.org/ontology**
- [DEF], TechnicalStaff, Thesis, Thing, Title, Topic, Undergraduate, **University**, Unpublished, With, Worked
- SemanticWebDocument, RDFXML, 2006-01-20, 43K, ontoRatio(1.00), metadata, cached

**http://www.aktors.org/ontology/portal**
- [DEF], Distance-Teaching-**University**, District, Edited, Edited-Book, Edited-Book-Reference, Editor, Edu
- SemanticWebDocument, RDFXML, 2004-12-10, 89K, ontoRatio(0.90), metadata, cached

**http://lsdis.cs.uga.edu/projects/semdis/opus**
- SemanticWebDocument, RDFXML, 2006-10-20, 21K, ontoRatio(1.00), metadata, cached
Document Search of Swoogle

ontology  document  term  across ontologies
Nanjing

Swoogle Search

list documents matching document search

http://destinfishingcharters.co/nanjing-electrical-area-of-trading-selection-2-

http://www.w3.org/2009/Talks/0829-Nanjing-IH/
SemanticWebDocument, RDFXML, 2011-11-01, 4K, metadata, cached

http://www.w3.org/2009/Talks/0830-Nanjing-IH/
SemanticWebDocument, RDFXML, 2011-11-08, 4K, metadata, cached
Term Search of Swoogle

list terms (URI references) matching term search

**range**
[DESC] "relationship is expected to point to a type "person" , so the "range" of "author" is "per"
[TYPE] [owl.DatatypeProperty=1, owl.ObjectProperty=9, owl.Property=1, rdf.Property=335, r
http://www.w3.org/2000/01/rdf-schema#range, metadata

**Person**
[TYPE] [owl.Class=5650, owl.Thing=1, rdfs.Class=178, rdfs.Resource=3]
http://xmlns.com/foaf/0.1/Person, metadata

**nick**
[DESC] ). ||| Defines the nickname of a person. ||| Nickname ||| testing
http://xmlns.com/foaf/0.1/nick, metadata

**knows**
[DESC] "A person known by this person (indicating some level of reciprocated interaction be
[TYPE] [owl.ObjectProperty=1039, owl.SymmetricProperty=1, owl.TransitiveProperty=1, rdf.I
http://xmlns.com/foaf/0.1/knows, metadata
Metadata of foaf:knows in Swoogle

Swoogle’s Metadata

- usesNamespace: http://xmlns.com/foaf/0.1/
- hasLocalname: knows

Cached Metadata

<table>
<thead>
<tr>
<th>NO</th>
<th>frequency</th>
<th>subject</th>
<th>predicate</th>
<th>?object</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1147</td>
<td><a href="http://xmlns.com/foaf/0.1/knows">http://xmlns.com/foaf/0.1/knows</a></td>
<td>rdf:type</td>
<td><a href="http://www.w3.org/2002/07/owl#ObjectProperty">http://www.w3.org/2002/07/owl#ObjectProperty</a></td>
</tr>
<tr>
<td>2</td>
<td>422</td>
<td><a href="http://xmlns.com/foaf/0.1/knows">http://xmlns.com/foaf/0.1/knows</a></td>
<td>rdfs:label</td>
<td>Foaf:knows</td>
</tr>
<tr>
<td>3</td>
<td>374</td>
<td><a href="http://xmlns.com/foaf/0.1/knows">http://xmlns.com/foaf/0.1/knows</a></td>
<td>rdfs:label</td>
<td>Relation:Knows</td>
</tr>
</tbody>
</table>
2007: SWSE (Harth et al., 2007)

- Keyword-based
- Pagerank-liked ranking
- Combining RDF graph ranks with data source ranks
- Filtering results by specifying a class
Semantic Search

2007: Watson (d’ Aquin et al., 2007)

- Organize resulting objects by documents / ontologies
- User can specify the searching scope
  - Local name
  - Labels
  - All literals
Searching Steve Jobs in Watson

1- http://lucas.is-a-geek.net/blog/2003/05/11/steve_jobs_quotes_hst.rdf
   ◦ http://lucas.is-a-geek.net/blog/2003/05/11/steve_jobs_quotes_hst
   ◦ http://www.mahalo.com/Steve_jobs
   ◦ http://www.answers.com/topic/steve-jobs
   ◦ http://kernelpanic.typepad.com/apple/2006/09/steve_schmidt_e.html
   ◦ http://www.tuaw.com/2006/03/20/steve-jobs-magazine-covers
   ◦ http://www.manzaneando.com/?p=62
   ◦ http://kromeblog.kromeboy.net/index.php/2007/09/06/100-con-le-scuse-di-steve
   ◦ http://www.techmeme.com/070906/p80 More...
Another Watson

- An open-domain QA system
- IBM DeepQA Project
  - MIT: Adaptive View-Based Appearance Model
  - UT: Reasoning and General Knowledge
  - USC: Information Extraction and Analysis
  - RPI: Virtualization Tools
  - University at Albany: Assurance of capacity for massive system
  - University of Trento: Self-study, Conversation
  - University of Massachusetts: Information Retrieval
  - Carnegie Mellon: QA basic algorithms
Another Watson

**Feature:**
- Unstructured text;
- 5 secs into more than 200 million pages;
- Understanding human language, subtle meaning, paddles, humor, satire...
- Determining the certainty of the answer
- No help from the WWW and engineers

**Beat ken and Brad on Jeopardy in 2011;**
• Future
  o Business Decision-Making
  o Medical Treatment
Semantic Search

- 2008: Sindice (Oren et al., 2008)
  - Property-value pair look-up knowing a property of an object
  - Keyword-based RDF document and Microformat search
  - Not focused on linked objects
Searching documents in Sindice

Sindice search: software found 1,915,998 documents (in 0.01 seconds)

**Software** (RDF)
- 2010-07-26 – 348 in 42.3 kB
  http://dbpedia.org/resource/Software (Search) Inspect: (Cache) (Live)

**software package** (RDF)
- 2010-08-25 – 129 in 31.6 kB
  http://sw.opencyc.org/2008/06/10/concept/en/SoftwarePackage (Search) Inspect: (Cache) (Live)

**software suite** (RDF)
- 2010-07-20 – 101 in 25.5 kB

**About: Free software** (RDF)
- 2010-04-21 – 167 in 48.1 kB
  http://dbpedia.org/page/Free_software (Search) Inspect: (Cache) (Live)
Searching documents using property-value pair in Sindice

Sindice search: * <foaf:mbox> "mailto:timbl@w3.org" found 24 documents (in 1.47 seconds)

"Tim Berners-Lee's FOAF file" (RDF)
+ 2011-05-23 – 78 in 10.0 kB
http://www.w3.org/People/Berners-Lee/card (Search) Inspect: (Cache) (Live)

"Tim Berners-Lee", "Timothy Berners-Lee" (RDF)
+ 2011-05-23 – 78 in 10.0 kB
http://www.w3.org/People/Berners-Lee/card#i (Search) Inspect: (Cache) (Live)

"Tim Berners-Lee's FOAF file" (RDF)
+ 2011-07-01 – 78 in 10.1 kB
http://www.w3.org/People/Berners-Lee/card.rdf (Search) Inspect: (Cache) (Live)

Tim Berners-Lee (RDF)
+ 2010-11-23 – 136 in 18.9 kB
http://semanticweb.org/id/Tim_Berners-2DLee (Search) Inspect: (Cache) (Live)
SPARQL Query in Sindice

```sparql
SELECT DISTINCT * WHERE {
  ?product a <http://purl.org/media/audio#Recording>.
}
```
Semantic Search

- 2009: Falcons (Cheng et al., 2009)
  - http://iws.seu.edu.cn/services/falcons/
  - Searching entities / linked objects / RDF documents
  - Keyword-based
  - Searching based on *Virtual Document*
  - Ranking results based on relevance and popularity
  - Class-based query refinement
“Who is the demo chair of ISWC 2008?”
query with “ISWC2008” AND “demo chair”
“Find relations between Chris Bizer and Tom Heath” query with “Chris Bizer” AND “Tom Heath”
Falcons – Scenario 2

After type refinement
Falcons – System Architecture
Falcons- Data Crawling

- Seeding
  - Searching filetype:rdf / filetype:owl in Google with keywords from Open Directory Project
  - Sampling from pingthesemanticweb.com / schemaweb.info / ...
  - Sampling from Linked Open Data Project

- Crawling
  - 300,000 RDF documents per day
  - 20M RDF documents total (55% well-formed) on August 2008
  - 600M quadruples
The distribution of the number of RDF documents on pay-level domains
• Virtual Documents of an object
  o local name
  o rdfs:label
  o rdfs:comment
  o other annotations
  o Virtual documents of neighbors
VD(eg1:Journal) = \alpha \cdot "journal" + \beta \cdot "article" + \theta \cdot "magazine"
• Query Relevance
  o Cosine similarity between query and virtual document

\[
\text{TextSim}(o, q) = \frac{V\text{Doc}(o) \cdot \text{Vector}(q)}{|V\text{Doc}(o)| \cdot |\text{Vector}(q)|},
\]

• Popularity
  o The occurrence of the object in the document set

\[
\text{Popularity}(o) = \log(|\text{Docs}(o)| + 1).
\]
Falcons – Other Features

- Query-relevant Snippet Generation
  - PD-Thread

- Query Refinement with Class Hierarchies
  - Filtering the resulting objects with class Restrictions
  - Discovering implicit typing information by class-inclusion reasoning
  - Recommending subclasses for incremental refinement


Semantic Search

Thanks