Towards Semantic Search on the Web: A Survey

Zhenning Shangguan

Apr. 14, 2009
Outline

• Introduction

• Existing Systems

• Technological Anatomy

• Conclusion
Introduction

• What is Semantic Search?
  – Users
    • Ordinary users
    • High-end users
  – Input
    • Keywords
    • Formal queries
    • Natural language questions
  – Output
    • URLs to (semantic) web documents
    • URIs to semantic entities
    • Snippet of KBs
  – Scale
    • Standalone applications
    • Intranet portals
    • Web
Introduction

• Where are the “semantics”?  
  – Ontologies  
    • Directly search for ontologies  
    • Keyword disambiguation  
    • Query formulation  
  – Query & store facilities  
    • Triple/quad store as repositories  
    • Formal queries to find entities/relations  
  – Reasoning  
    • Identity consolidation using IFPs  
    • Class subsumption reasoning to help navigation  
    • Web-scale reasoning to find hidden knowledge
“Semantic Search” Systems

Can we arrive here?
PowerSet

• Enable ordinary people to interact with information naturally and effectively (Pell ’07)
• Combine deep NLP and scalable search technology
  – Interpret the web
  – Index the interpreted web
  – Interpret the queries
  – Search by matching
• Information extracted from Wikipedia, and search results integrated from various sources (e.g., FreeBase)
Swoogle

• Semantic web documents & metadata search engine for high-end users (Li et al. ’04)
• Crawler-based IR system on the semantic web
  – Crawls the semantic web (SWDs & SWTs)
  – Indexing (traditional IR techniques)
  – Ranking (OntoRank based on Rational Surfer Model)
  – Querying using keywords
• Approx 2.9 million SWDs indexed
Falcons

- Semantic web objects & concepts (RDF mainly) search engine for high-end users (Cheng ’06 ’08)
- Mostly like Swoogle, with highlighting features
  - Class hierarchy based navigation among ontologies (class subsumption reasoning)
  - Comprehensive text descriptions of semantic web objects (RDF sentence) (Zhang ’07)
  - Semantic web object summarization
- Approx 10 million SWDs indexed
SearchMonkey

• Leverage structured metadata to enrich presentations of search results (Mika ’08)
• Aim to create an ecosystem of publishers, developers and end-users
  – Publishers generate structured metadata
    • Standard SW technologies, e.g., RDFa, SIOC, etc
  – Developers create apps using these metadata
  – End-users enhance searching experience
• Web-scale platform, integrated tightly with the core search engine of Yahoo!
Technological Anatomy

• Components of a web search engine (Brin ’98)
  – Crawling
  – Indexing
  – Storage
  – Ranking

• What about the search applications on the Semantic Web?
Crawling

• Different kinds of documents exist on the Semantic Web
• Different popularities of different types of documents
• Where to find an efficient and scalable crawler that can best harvest knowledge from the Semantic Web?
Crawling

- **Slug (or Scutter) (Dodds ’03)**
  - Configurable, flexible, and modular framework
  - Implemented in Java using Jena API
  - RDF vocabulary for describing configuration & activity
- **Hybrid Semantic Web Harvesting Framework (Li ’06)**
  - Integration of several harvesting frameworks
    - Meta crawling
    - Bounded HTML crawling
    - RDF crawling
  - Starting seed URLs provided via manual submission
  - Harvested SWDs as training data to generate new seed URLs
Crawling

• MultiCrawler (Harth ’06)
  – Hybrid crawling
    • Semantic documents
    • Web pages with free texts
  – Crawling & indexing contains 5 steps
    • Fetching, detecting, transforming, indexing, extracting
    • Textual information also transformed and indexed
  – Parallelized and distributed crawling

• Semantic Sitemap (Cyganiak ’07)
  – Extension to the sitemap protocol
  – State where the RDF data is, and alternative methods to access it (dumps, SPARQL endpoints)
Indexing

• Inverted Indexing
  – Language processing based on variations of bag-of-words model or probabilistic linguistic model
  – VSM model to calculate term-document relevance
  – Swoogle (Li et al ’04), Semplore (Zhang et al ’07)

• Refined inverted indexing
  – Pure IR style indexing faces the problem of ambiguity
  – SENSE: inverted index for word meaning (Basile ’08)
    • Lexical level
    • Word meaning level – wordnet
    • Named entity level – named entity recognition
Storage

• SWSE uses YARS as quadruple repository (Harth ’08)
• Index quadruples to store provenance information with triples
• Two sets of indexes
  – Lexicon: covers string representation of RDF nodes, enables fast retrieval of RDF nodes
  – Quad index: covers access patterns of RDF quadruples
Ranking

• Most of the existing ranking schemes are variations of PageRank
  – Swoogle (Li et al ’05)
    • Based on Rational Surfer Model
    • Three levels of granularity: docs, terms, RDF sub-graphs
  – ReConRank (Hogan et al ’06)
    • Resource Rank: ranking on focused sub-graph (Kleinberg ’99)
    • Context Rank: logically treating context as resource nodes, and generate implicit links for ranking
Query Formulation

• Domain ontology based interpretation
  – High precision, low recall solutions
  – No extensibility, domain dependent
  – No portability, dedicated to certain platform
  – Examples:
    • Generate formal queries using templates (Lei ’06)
    • Translate keywords to conjunctive formal queries (Tran ’08)

• Query ranking
  – Based on sub-graph matching algorithms
  – For performance, indexing sometimes needed
  – Balance between precision and recall
  – Examples:
    • SPARK (Zhou ’08)
    • Q2Semantics (Wang ’08)
Reasoning

• Controversial topic
  – Computational expenses vs. extra credits?
• Some reasoning are employed
  – Identity consolidation in SWSE (Hogan ’08)
  – Class subsumption reasoning in Falcons (Cheng ’08)
  – MaRVIN: web-scale online massive inference (Oren ’08)
• Some attempts to integrate reasoning with search
  – IR on the web should be automated by binding search and inference together (Mayfield & Finin ’03)
  – Reasoning with limited rationality is truly rational on the web scale (Fensel & Harmelen ’07)
Conclusion

• No clear definition of semantic search exists
  – Search for semantic documents on the web
    • Mixture with traditional web search (due to current maturity of SW)
  – Search with the help of semantic technologies
    • Merging trend of enhancing IR with semantics, e.g., SearchMonkey (Mika ’07)
  – Search for knowledge on the web
    • NAGA (Weikum ’08)
      – Integration of IR and DB to build world-wide KB to support knowledge search
Conclusion

• Techniques from IR and other communities widely adopted
  – Simply treat semantic documents and entities as web documents and words
  – Many semantic search engines use inverted index
  – Searching based on string matching techniques
  – Ranking based on graph link analysis
    • But how can we do KnowledgeRank on the global web?
Conclusion

• Benefits of semantic technologies have not shown explicitly
  – Searching for semantic web documents might only benefit knowledge engineers
    • Ordinary users cannot feel the semantic values directly
  – Semantics are often used during query formulation
    • computationally expensive and domain dependent
  – Web scale knowledge extraction and integration is still challenging (Weikum ’08, Decker ’08)
    • Ongoing efforts of automated ontology learning using NLP and ML, but with limited range and partial success
  – Some researchers believe semantics helps in automating the traditional searching procedure
    • keyword → list of pages → keyword refinement → search again
Conclusion

- Scalability may be a big issue
  - Engineering problem vs. Research problem
  - Current repositories and query languages does not scale to “frillions” of triples
  - If reasoning is favored, needs to sacrifice soundness and completeness for performance

- Often lack of convincing evaluation
  - Hard to evaluate systems running in web scale
  - Synthesized data may not reflect real world
Questions

• Thanks