의미기반 검색엔진의 동향과 핵심 기술

한 성 국

의미기술 연구소
원광대학교
Agenda

1. Overview of the current Search technology
2. The State-of-the-art of Semantic Search
3. Core Technologies of Semantic Search
4. Summary
Google by far comprises the largest share of searches.

- Microsoft has been trying to buy Yahoo to increase Microsoft’s search share. As of June 12th, both companies have ended merger talks.
- Now, Microsoft merges Powerset...
Rich Content and Vertical Search

Video: http://kr.youtube.com/

Map: http://maps.live.com/

Blog: http://www.google.com/blogsearch

People: http://www.pipl.com/

www.stskorea.org
Rich Content and Vertical Search

- Amazon
- Articles
- Wikipedia

- Blogs
- Photos
- Flickr

- del.icio.us
- Events
- Upcoming.org

- Book marks

- Last.fm
- Places
- Dopplr

- Music

- Movies
- Products
- Microsoft Aura

www.stskorea.org
User-Friendly Interface

Tree

http://www.tafiti.com/

Network

http://www.kartoo.com/

Space

http://www.quintura.com/
# Web 2.0 and Search

## Applications
- Wiki, Blog, Social Network, UCC, Podcasting, RSS,...

## Realization Technologies
- Mashup, Folksonomy, AJAX, Microformat, Open API, SOAP(REST),...

## Principles
- The Web As Platform
- Harnessing Collective Intelligence
- Data is the Next Intel Inside
- End of the Software Release Cycle
- Lightweight Programming Models
- Software Above the Level of a Single Device
- Rich User Experiences

## Ethos of Web 2.0
- Openness/Sharing
- Participation/Collaboration

## Foundational Basis
- Long Tail
- Collective Intelligence
- Data on a epic scale
- Services, not packaged software
- Reuse

## Tag Cloud
- art
- architecture
- animal
- api
- architecture
- art
- australia
- baby
- barnacle
- beach
- berlin
- birthday
- box
- black
- blue
- blossom
- bw
- california
- cameraphone
- camping
- canada
- car
- cat
- chicago
- china
- christmas
- church
- city
- clouds
- cloud
- concert
- day
- dog
- dogs
- england
- europe
- family
- florida
- flower
- flowers
- food
- france
- friends
- fun
- garden
- geotagged
- germany
- get
- graffiti
- green
- halloween
- hawaii
- sky
- holiday
- home
- house
- india
- japan
- italy
- japan
- lake
- landscape
- light
- london
- losangeles
- macro
- march
- may
- me
- mexico
- nativity
- mountains
- museum
- music
- nature
- new
- newyork
- nyfc
- ocean
- paris
- park
- party
- people
- portrait
- red
- rover
- redrover
- rock
- san
- scottish
- sea
- seattle
- sky
- snow
- spain
- spring
- street
- summer
- sunset
- sydney
- taiwan
- town
- thailand
- tokyo
- toronto
- travel
- tree
- tree
- trip
- uk
- urban
- usa
- vacation
- vancouver
- washington
- water
- wedding
- white
- winter
- yellow
- zoo

## Folksonomy
- circle
- green
- hexagon
- blue

## Web 2.0 and Search

## Tag Cloud
- art
- architecture
- animal
- api
- architecture
- art
- australia
- baby
- barnacle
- beach
- berlin
- birthday
- box
- black
- blue
- blossom
- bw
- california
- cameraphone
- camping
- canada
- car
- cat
- chicago
- china
- christmas
- church
- city
- clouds
- cloud
- concert
- day
- dog
- dogs
- england
- europe
- family
- florida
- flower
- flowers
- food
- france
- friends
- fun
- garden
- geotagged
- germany
- get
- graffiti
- green
- halloween
- hawaii
- sky
- holiday
- home
- house
- india
- japan
- italy
- japan
- lake
- landscape
- light
- london
- losangeles
- macro
- march
- may
- me
- mexico
- nativity
- mountains
- museum
- music
- nature
- new
- newyork
- nyfc
- ocean
- paris
- park
- party
- people
- portrait
- red
- rover
- redrover
- rock
- san
- scottish
- sea
- seattle
- sky
- snow
- spain
- spring
- street
- summer
- sunset
- sydney
- taiwan
- town
- thailand
- tokyo
- toronto
- travel
- tree
- tree
- trip
- uk
- urban
- usa
- vacation
- vancouver
- washington
- water
- wedding
- white
- winter
- yellow
- zoo

## Folksonomy
- circle
- green!
- hexagon?
- blue?
Marketing with Search Engine

Rank! Rank!! Rank!!!

Search Engine Optimization (SEO) and Search Engine Marketing (SEM)

1. Page Titles
2. URLs
3. Anchor Text
4. Website Architecture
5. Link Title & Alt Images
6. Relevant content
7. Sitemap.xml
8. Hosting
9. Freshness

www.stskorea.org
Beyond the Limits of Keyword Search

The World Wide Web
- Web 1.0 (1990 - 2000)
  - The Desktop
  - Databases
  - Files & Folders
  - Directories
- Keyword search

Web 2.0 (2000 - 2010)
- The Social Web
  - Social networking
- Tagging

Web 3.0 (2010 - 2020)
- The Semantic Web
  - Semantic search
- Reasoning

Web 4.0 (2020 - 2030)
- The Intelligent Web
  - Natural language search
  - Semantic search
  - Reasoning

Amount of data
Productivity of Search
- No definitive formulation.
- Considerable uncertainty.
- Complex interdependencies.
- Incomplete, contradictory, and changing requirements.
- Stakeholders have radically different world views and different frames for understanding information.
The Age of Semantic Search

Yahoo makes semantic search shift

Powerset Previews Semantic Search Engine

Powerset aims to prove that its natural language-based search makes it easier to find things on Wikipedia and Freebase, an open community database.

By Thomas Claburn

Microsoft buys Powerset, gets foot in semantic search door

By Jacqui Cheng | Published: July 01, 2008 - 05:21PM CT

While Google and Yahoo focus on indexing Flash sites, Microsoft has other search plans up its sleeve. As the rumors have foretold, Microsoft announced today that it has acquired Powerset and plans to integrate the company’s search and natural language features into Live Search. The San Francisco-based company’s employees will remain in place, but will become part of Microsoft’s Search Relevance team.
Hakia: semantic search... set to music
By Nate Anderson | Published: July 30, 2007 - 11:26PM CT

The intersection of search and samba

If you ever wondered what search queries sound like when set to music, Hakia has your answer. Employees at the beta semantic search engine have put together a band and recorded an album of songs based on actual user queries, things like "Weapons of Mass Instruction." Think samba meets lounge music meets beat poetry and you get the idea. It's a little... odd.

But Hakia doesn't mind being odd. The company was born out of a desire to do search differently. Search engines generally don't understand either content on the Web or the content of user queries; they work through keyword analysis, link weighting, and other statistical methods that allow an engine to produce results ever needing to understand the implicit question in the query. Hakia is about semantic understanding. It provides traditional search capabilities, but Hakia executives can provide the results talked with Hakia's CEO, Pulatkonak, the company stake out territory in a province already claimed by Google, Yahoo! and Microsoft.

IBM Unveils Semantic Search For Corporate E-mail

The OmniFind Personal E-mail Search aims to make associations between the underlying concepts of words often used in corporate e-mail.

By Antone Gonsalves
InformationWeek
12/20, 2007 04:57 $?

IBM (NYSE: IBM) on Thursday introduced a test version of a semantic search engine for companies that want to give employees more advanced tools for searching their e-mail stored in Microsoft (NSDQ: MSFT) Outlook or IBM's Lotus Notes.

The OmniFind Personal Email Search, developed by researchers at IBM facilities in California, Israel, and India, takes search beyond keywords by being able to make associations between the underlying concepts of words often used in corporate e-mail. In doing that, IBM is looking to offer customers technology that can help them retrieve useful information hidden in e-mail databases.
Semantic Search attempts to augment and improve traditional search results by using data from the SW.

<table>
<thead>
<tr>
<th></th>
<th>Syntactic Search</th>
<th>Semantic Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document View</td>
<td>Bag-of-Words</td>
<td>Vocabularies and Concepts</td>
</tr>
<tr>
<td>Search Approach</td>
<td>Word matching</td>
<td>Concept matching</td>
</tr>
<tr>
<td>Search Process</td>
<td>One hot</td>
<td>Reasoning / Inference</td>
</tr>
</tbody>
</table>

Ontology and Semantic Search

- Help user formulate semantic queries
- Re-formulate or re-interpret queries
- Browse domain
- Formulate related queries
- Interoperability between search application
- Semantic indexing of documents
## Typical Semantic Search Engine

<table>
<thead>
<tr>
<th>Search Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Search</td>
<td>Freebase, Yahoo! Microsearch, ...</td>
</tr>
<tr>
<td>Natural Language Search</td>
<td>Powerset, Hakia, AskMeNow, AskWiki, ...</td>
</tr>
<tr>
<td>Vertical Search</td>
<td>Kango, ...now UpTake, AdaptiveBlue, ...</td>
</tr>
<tr>
<td>Social Networking Search</td>
<td>SemantiNet, Delver, Google Social Graph API,</td>
</tr>
<tr>
<td>Personalized Search</td>
<td>Twine, MavinIT, PSS, ...</td>
</tr>
</tbody>
</table>
Freebase

- MetaWeb Technologies
- Covers millions of topics in hundreds of categories from large open repositories like Wikipedia, MusicBrainz, and the SEC archives.
- Topics are organized by type. Type are grouped together into domains.
- A particular topic exists only once, even if it might normally be found in multiple databases to reconcile many facets of one topic.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Jump to type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts &amp; Entertainment</td>
<td>Science &amp; Technology</td>
</tr>
<tr>
<td>Society</td>
<td>Time &amp; Space</td>
</tr>
<tr>
<td>Sports</td>
<td>Special Interests</td>
</tr>
<tr>
<td>Products &amp; Services</td>
<td>System</td>
</tr>
<tr>
<td>Money</td>
<td></td>
</tr>
</tbody>
</table>

**Data Mob Projects**

**Olympics are on!**
Now the Olympics are on, our main data mob task will be to keep Freebase up to date with...

**Olympic Games: teams for 2008**
Here’s a list of Olympic teams in this year’s games. We’ve loaded the entire US team...

Recent Discussions about Domains and Types
Powerset builds a transformative consumer search engine based on Natural Language Processing (NLP).

Powerset reads and understands every sentence on a Webpage and allows asking questions in plain English.

PowerMouse is natural Language Indexer to show in dramatic fashion how compactly large amounts of data can be organized and displayed based on a few semantic relationships.
UpTake

- Vertical semantic search engine for personalized travel information.
- UpTake indexes the collective wisdom on travel from the entire Web.
- UpTake is creating an ontology of global travel content that includes ranking of superlatives within review sites.
- UpTake returns results according to user preference.
SemantiNet

- SemantiNet leverages Semantic Web concepts to seamlessly integrate information and services.
- SemantiNet collects relevant information from common social networks, and established Web sites in order to provide users with a customized and efficient personalized and contextual browsing experience.
Twine

- Radar Networks,
- Twine is a "knowledge networking" tool designated as a revolutionary Semantic Web application.
- Twine enables individuals and groups to organize, share and discover information and knowledge around their interests.
- It has aspects of social networking, wikis, blogging, knowledge management systems – but its defining feature is that it's built with Semantic Web technologies.
- Semantic Web technologies are being used: RDF, OWL, SPARQL, XSL, GRDDL.
University of Maryland Baltimore County

- a crawler-based indexing and retrieval system for the Semantic Web. It extracts metadata for each discovered document, and computes relations.
  - Finding appropriate ontologies by their popularity
  - Finding instance data
  - Characterizing the Semantic Web to reveals interesting structural properties such as “how the Semantic Web is connected”, “how ontologies are referenced”, and “how an ontology is modified externally”.

Swoogle Today

<table>
<thead>
<tr>
<th>admin_dt</th>
<th>2008-08-21 00:19:49</th>
<th>Datetime Watched</th>
</tr>
</thead>
<tbody>
<tr>
<td>url_total</td>
<td>7,760,194</td>
<td>Number of URLs being discovered</td>
</tr>
<tr>
<td>url_pinged</td>
<td>4,227,882</td>
<td>Number of URLs being pingered</td>
</tr>
<tr>
<td>total_swd</td>
<td>2,629,354</td>
<td>Number of Semantic Web Documents (regardless of embedded or containing some errors) be confirmed</td>
</tr>
<tr>
<td>total_swd_strict</td>
<td>1,478,172</td>
<td>Number of error-free pure Semantic Web Documents</td>
</tr>
<tr>
<td>total_swd Embed</td>
<td>930,393</td>
<td>Number of documents (except SWDs, PDF, and JPEG) embedding Semantic Web Data</td>
</tr>
<tr>
<td>triple_total</td>
<td>661,810,023</td>
<td>Number of triples could be parsed from all Semantic Web Documents</td>
</tr>
</tbody>
</table>

www.stskorea.org
5 Core technologies for Semantic Search

- Tagging
- Statistics
- Linguistics
  - Natural language Processing
- Semantic Web
  - Metadata / Ontology
- Artificial Intelligence

User Interaction

Concept organization

Reasoning

Service Architecture

- Architecture
- Concept coupling
- User Interaction
- User Context
- Query Construction
- Query Modification
- Ontology
## Categorical Features of Semantic Search Engine

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Stand-alone</th>
<th>Maintain an concept index of document</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Meta Search</td>
<td>Use subordinate search engines</td>
</tr>
<tr>
<td>Coupling between documents and ontologies</td>
<td>Tight coupling</td>
<td>Data of documents refer explicitly to concepts of a specific ontology.</td>
</tr>
<tr>
<td></td>
<td>Loose coupling</td>
<td>Not committed to any available ontology</td>
</tr>
<tr>
<td>User Interaction</td>
<td>Transparent</td>
<td>Semantic capabilities invisible to the user.</td>
</tr>
<tr>
<td></td>
<td>Interactive</td>
<td>Ask for clarification or recommendation</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>Both</td>
</tr>
</tbody>
</table>
### Categorical Features of Semantic Search Engine

<table>
<thead>
<tr>
<th>User context</th>
<th>Learning</th>
<th>Extract from user interaction dynamically</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard-coded</td>
<td>Ask for query category</td>
<td></td>
</tr>
<tr>
<td>Query modification</td>
<td>Manually</td>
<td>The user modifies a query.</td>
</tr>
<tr>
<td></td>
<td>Query rewritten</td>
<td>A query can be optimized by the system.</td>
</tr>
<tr>
<td></td>
<td>Graph-based</td>
<td>Use graph traversal algorithm</td>
</tr>
<tr>
<td>Ontology Construction</td>
<td>anonymous</td>
<td>Disregard the vocabulary and the semantics</td>
</tr>
<tr>
<td></td>
<td>Standard property</td>
<td>Synonym, hyponym,...</td>
</tr>
<tr>
<td></td>
<td>Domain-specific property</td>
<td>Domain ontology</td>
</tr>
<tr>
<td>Ontology technology</td>
<td>Language</td>
<td>RDF, OWL,...</td>
</tr>
</tbody>
</table>
# Analysis of Semantic Search Technology

<table>
<thead>
<tr>
<th>Search phase</th>
<th>Feature</th>
<th>Functionality</th>
<th>Interface Components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Free text input</td>
<td>• keyword(s)</td>
<td>• Single text entry</td>
</tr>
<tr>
<td>Query construction</td>
<td></td>
<td>• natural language</td>
<td>• Property-specific fields</td>
</tr>
<tr>
<td></td>
<td>Operators</td>
<td>• Boolean operators</td>
<td>• Application-specific syntax</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• special purpose operators</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• regular expressions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Controlled terms</td>
<td>• disambiguate input</td>
<td>• Value list</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• restrict output</td>
<td>• Faceted</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• select predefined queries</td>
<td>• Graph</td>
</tr>
<tr>
<td></td>
<td>User feedback</td>
<td>• pre-query disambiguation</td>
<td>• Suggestion list</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Semantic auto completion</td>
</tr>
<tr>
<td>Search algorithm</td>
<td>Syntactic matching</td>
<td>• exact, prefix, substring match</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• minimal edit distance</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• stemming</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semantic matching</td>
<td>• thesauri expansion</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• graph traversal</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• RDFS/OWL reasoning</td>
<td></td>
</tr>
</tbody>
</table>
## Analysis of Semantic Search Technology

<table>
<thead>
<tr>
<th>Search phase</th>
<th>Feature</th>
<th>Functionality</th>
<th>Interface Components</th>
</tr>
</thead>
</table>
| Query construction | Data selection       | • Selected property values  
                    • Class specific template  
                    • Display vocabulary | • Text  
                    • Graph  
                    • Tag cloud  
                    • Map  
                    • Timeline  
                    • Calendar |
|                    | Ordering             | • Content and link structure based ranking                          | • Ordered list                           |
|                    | Organization         | • Clustering by property or path  
                    • Dynamic clustering                                                  | • Tree  
                    • Nested box structure  
                    • Cluster map                                                      |
|                    | User feedback        | • Post-query disambiguation  
                    • Query refinement  
                    • Recommendation of related resources                              | • Facets  
                    • Tag cloud  
                    • Value list                                                    |

Semantic Search is a kind of Generic tasks.
- More than simple document search
- Diverse applications in BioInformatics, EcoScience, Medical Science...

Ontology is a key player of Semantic Search.
- RDFa, Microformat, GRDDL,...
- RDF, RDF Schema, OWL,...
- Ontology Annotation and Population
- SPARQL and Query processing,

Multi-disciplinary research and development.
- Natural Language Processing and Text Mining
- Web Science

User-friendly
- Diverse vertical semantic search with domain ontologies
- Visualization
- Mobile Search
의미기반 검색엔진의 동향과 핵심 기술

경청해 주시어, 감사드립니다.

한 성 국
skhan@wku.ac.kr
의미기술 연구소
원광대학교