Semantic Search via XML Fragments: A High-Precision Approach to IR

Jennifer Chu-Carroll
IBM T.J. Watson Research Center

April 2006

Joint work w/ John Prager, Krzysztof Czuba, David Ferrucci & Pablo Duboue
Overview

- Why Semantic Search?
- Corpus Analysis for Semantic Search
- Using XML Fragments for Semantic Search
  - Three XML Fragments operators
  - Applied to address four query-time semantic needs
- Evaluation
- Pseudo demo
- Conclusions
Why Semantic Search?

- **Plenty of people find what they need with keywords**
  - Web search
  - Desktop search

- **These applications work well with keywords because**
  - Web search: corpus is huge
  - Desktop search: looking for known item

- **Keyword search does not work quite so well for**
  - Enterprise search: limited-size corpus
  - Exploratory research: user is researching a topic of interest and does not know what they are looking for

- **Semantic search can help in the latter classes of applications**
Semantic Search vs. Keyword Search

SEARCH: Going rate for leasing a billboard near Triborough Bridge

Top hits from popular search engines miss the mark…

*Keywords may match*

*BUT WRONG content returned*

*And right content MISSED*
SEARCH: Going rate for leasing a billboard near Triborough Bridge

Wired New York Forum - Long Island City Development
"it's Right over The 59th Street Bridge." ... it's going to become our SoHo. The Powerhouse is most likely going to become a performing Arts facility for ... www.wirednewyork.com/forum/printthread.php?t=5160&pp=50 - 150k - Supplemental Result - Cached - Similar pages

Long Island City Development [Archive] - Wired New York Forum
... Creek to the Triborough Bridge - would become an attractive destination ... Several Long Island City sites are now in play - Junger himself’s going ... www.wirednewyork.com/forum/archive/index.php/t-5160.html - 255k - Cached - Similar pages
[ More results from www.wirednewyork.com ]

[PDF] Austin’s Wireless Future
File Format: PDF/Adobe Acrobat - View as HTML
In Section 2 we explore the near-term future of wireless—ubiquitous mobile ... applications such as the Triborough Bridge installation, wireless sensors ... www.ic2.org/publications/AustinsWirelessFuture.pdf - Similar pages

[PDF] Funding of the High-Speed Rail System
File Format: PDF/Adobe Acrobat - View as HTML
Triborough Bridge and Tunnel Authority* - Oklahoma, Turnpike Authority ... management fees, reserves for going TIs, and leasing commissions. ... www.cahighspeedrail.ca.gov/plan/pdf/Financial_Plan.pdf - Similar pages

The rail magazine for The computer age ...
Surprisingly, thanks to 9600 baud modems, lower rates, and automated access ... his previous positions included President, Triborough Bridge and tunnel ... www.railfan.net/allpix/text/fbro_1_5.bit - Supplemental Result - Similar pages

Checklist 1969
Triborough Bridge and Tunnel Authority, selected procurement and contract ... Distribution of high school graduates and college going rate, New York State, ... www.mist/nyed.gov/edocs/education/check89.htm - 513k - Cached - Similar pages

[PDF] www.icc.utexas.edu/publications/AustinsWirelessFut...
File Format: PDF/Adobe Acrobat - View as HTML
Supplemental Result - Similar pages
**SEARCH:** Going rate for leasing a billboard near Triborough Bridge

* Were you looking for *Going rate for leaving a billboard near Triborough Bridge*

**Web Results**

Page 1 of 4 results containing *Going rate for leasing a billboard near Triborough Bridge* (0.70 seconds)

**Wired New York Forum - Long Island City Development**

... on Vernon Boulevard near the Queensboro Bridge, the founders of Silvercup ... Newtown Creek in the Triborough Bridge - would become an attractive ... Pearson St. - where he's going to complete a 20-story ...

**Cached page**

**FINANCIAL PLAN**

Triborough Bridge and Tunnel Authority** Oklahoma Turnpike Authority ** Average Annual Growth Rate** 1979-1982 10.9% 1980-1990 5 ... Leveraged Leasing of Rail Rolling Stock When rail rolling ...

**Cached page**  
**PDF file**

**NEW YORK: Stadiums/Arenas (NETS, JETS, YANKS, and others) [Archive ...**

... backward from the adjacent Triborough Bridge elevated roadway that leads to ... I was going to the name of one of my companies on it to get ... The area near the Continental Arena has become construction ...

**Cached page**

**http://odur.let.rug.nl/ftp/pub/prolog-app/DutchBrillTagging/TaggerSoftware/Bin_and_Data/LEXICON.WSJ**

Billboard NNP Billboarding NNP Billed VBN Billerica NNP Billheimer NNP Billiards NNP NN Billie NNP Billing NN Billings NNS NNP Billion CD Billionaire NN Billions NNS Billmeyer NNP Billock NNP Bills NNS ...


**Another Search Engine**

* Some Different* Hits but Still All Misses

Makes you wonder...

Could it be out there?
Remarkably…With some **Location** semantics

We can quickly find Hi-Res examples of area of interest

But NOT the information we need
Semantic Search Can Improve Recall

SEARCH: Going rate for leasing a billboard near Triborough Bridge

No Keywords in Common
But a good “hit”

“…We were offered $250,000/year in 2001 for an outdoor sign in Hunts Point overlooking the Bruckner expressway. …”
Semantic Search Can Improve Precision

SEARCH: Going rate for leasing a billboard near Triborough Bridge

Common Keywords
Bad Semantic Match

“…Simon and Garfunkel's "The 59th Street Bridge Song" was rated highly by the Billboard magazine in the 60's…”
Semantic Search on Semantically-Encoded Corpora

- Corpus: XML documents
  - `<Book>`
    - `<Title>`My Life`</Title>`
    - `<Author>`
      - `<FirstName>`Bill`</FirstName>`
      - `<LastName>`Clinton`</LastName>`
    - `<Publisher>`Knopf`</Publisher>`
    - `<PubDate>`2004`</PubDate>`
  - `<Book>`
    - `<Title>`The Survivor: Bill Clinton in the White House`</Title>`
    - `<Author>`
      - `<FirstName>`John`</FirstName>`
      - `<LastName>`Harris`</LastName>`
    - `<Publisher>`Random House`</Publisher>`
    - `<PubDate>`2005`</PubDate>`

- **Semantics enables more precise queries**
  - `<Book>`Bill Clinton`</Book>` retrieves both documents
  - `<Book>`
    - `<Author>`Bill Clinton`</Author>`
  - `<Book>`
    - retrieves only the first document

- **XML document corpora**
  - Automatically generated from databases
  - Manually marked up
  - From the semantic web (vision)

- **But… most existing electronic documents are unannotated or sparsely annotated**
Corpus Analysis for Semantic Search

- **Apply semantic search to previously unannotated documents**
  - Use existing text analysis techniques for document processing
    * Named entity recognition: PERSON, ORGANIZATION, DATE,…
    * Relationship extraction: ALIAS, OWNER, EXPORTSTO,…
  - Annotate and index documents with extracted semantic information
  - Leverage automatic annotations for semantic search

- **Example**

  `<BirthPlaceOf> <BirthDateOf> <Alias> <Person> President Clinton </Person> </Alias> on <Date> August 19, 1946 </Date> </BirthDateOf>, in <City> Hope, Arkansas </City> </BirthPlaceOf>, three months after his father died in a traffic accident.`
XML Fragments for Semantic Search

- **XML Fragments query language**
  - Keyword queries augmented with XML tags
    - `<CeoOf> <Person> Center </Person> </CeoOf>`
  - Supports classic query operators
    - Phrase: applies to content within XML tags
      - `<Person> “Thomas Ryan” </Person>`
    - “+” and “-”: apply to content within XML tags or XML fragments
      - `<Organization> +Clinton Institute </Organization>`
      - `+<Organization> +Clinton Institute </Organization> -Hillary`
  - Query language specifies valid syntactic representations but are semanticless

- **JuruXML indexer and search engine**
  - Supports indexing of keywords and annotations
  - Supports XML Fragments query language
  - Developed at IBM Haifa Lab
Three XML Fragment Operators

- **Conceptualization**
  - Generalizes a lexical string to a concept represented by that string
  - “animal” vs. “<Animal></Animal>”

- **Restriction**
  - Constrains the XML tags in which keywords in relevant documents can appear
  - “bass” vs. “<Animal> bass </Animal>”

- **Relation**
  - Specifies relations between terms that appear in relevant documents
  - Syntactic: “<SubjectVerb> Unabomber kill </SubjectVerb>”
  - Semantic: “<Kill> Unabomber <Person></Person> </Kill>”
  - Pragmatic: “<HasNegOpinion> Clinton war Iraq </HasNegOpinion>”
Application of XML Fragment Operators

- **XML Fragments operators**
  - Enhance the expressiveness of search queries
  - Should lead to more relevant search results

- **Operators applied to address four query-time semantic needs**
  - Specify target information
  - Disambiguate keywords
  - Specify search term context
  - Specify relations between search terms
Target Information Specification

- Application of the **conceptualization** operator
- Uses XML tags to represent target information as concepts
- E.g., user wants to know the zip code of the White House
  - Keyword query: +“white house” +zip +code
  - Semantic query: +“white house” +<Zipcode></Zipcode>

- **Useful when**
  - Query consists primarily of low idf terms; and
  - Target information does not frequently co-occur with other search terms

- Adopted in PIQUANT factoid QA system to represent answer types
Search Term Disambiguation

- Application of the **restriction** operator
- Uses XML tags for disambiguation of keywords
- E.g., User wants to know George Washington’s birthday
  - Query w/o disambiguation: `+<Date></Date> bear +George +Washington`
  - Retrieves: “A bear was spotted on the campus of George Washington University yesterday”
  - Query w/ disambiguation: `+<Date></Date> bear +<Person> +George +Washington </Person>`

- **Useful when**
  - Query term has multiple word senses in the corpus; and
  - User is interested in minority word sense

- **Adopted in**
  - PIQUANT QA for automatic disambiguation
  - SAW (Semantic Analysis Workbench) for interactive query refinement
Search Term Context Specification

- Application of the **restriction** operator
- Uses XML tags to specify the context in which query terms should appear
  - subject/object
  - agent/patient
  - request/suggest/opinion
- E.g., User wants to know what was the Pentagon panel’s position was with respect to the dispute over the US Navy training range in the island of Vieques
  - Keyword query: +Pentagon panel +US +Navy training range island +Vieques
  - Query w/ context: +Pentagon panel +<Quotation> +US +Navy training range island +Vieques </Quotation>

- Adopted in AQUAINT 2005 opinion pilot
Relation Specification

- Application of the **relation** operator
- Uses XML tags to specify relations between query terms
- E.g., User wants to know what biological weapons Iraq possesses
  - Query w/o relation: `<Iraq><BiologicalWeapon></BiologicalWeapon> possess`  
  - Matches: “Iraq possesses the technology to produce smallpox.”
  - Relation query: `<Owner> <Iraq <BiologicalWeapon></BiologicalWeapon></Owner>`  
  - Matches: “Iraq’s anthrax stockpile is estimated to be at least 8500 liters.”

- Adopted in SAW for automatic relation query generation and interactive query refinement
Evaluation

- **Corpora and corpora analysis**
  - AQUAINT corpus (3GB; 1M+ documents)
    - Annotated with ~100 named entity types
  - CNS corpus (~38,000 documents)
    - Annotated with ~100 named entity types and ~15 relation types

- **Standard test set and judgments used when possible**

- **In all experimental setups**
  - Application of XML Fragments operator constrains baseline query
  - Hypothesis: our applications of semantic search yield significant improvement in the precision of retrieved results

- **Evaluation metrics**
  - R-Prec: Precision at R (R is # of known relevant documents)
  - MAP: Mean Average Precision
  - Exact Precision
Target Information Specification Evaluation

- **Experimental setup:**
  - Test set: 50 factoid questions from TREC 2005 QA document task with relevant document judgments
  - Baseline query: keywords extracted from question
    - +White +House zip code
  - Query with target: keywords + answer type
    - +White +House zip code +<Zipcode></Zipcode>

- **Evaluation Results**

<table>
<thead>
<tr>
<th></th>
<th>R-Prec</th>
<th>MAP</th>
<th>Exact Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Query</td>
<td>0.4219</td>
<td>0.4329</td>
<td>0.0817</td>
</tr>
<tr>
<td>Query w/ Target</td>
<td>0.4342</td>
<td>0.4505</td>
<td>0.1124</td>
</tr>
</tbody>
</table>
Search Term Disambiguation

- **Experimental setup:**
  - Test set: 50 factoid questions from TREC 2005 QA document task with relevant document judgments
  - Baseline query:
    - +White +House zip code +<Zipcode></Zipcode>
  - Query with disambiguation: select keywords disambiguated
    - +<Facility>+White +House</Facility> zip code +<Zipcode></Zipcode>

- **Evaluation Results**
  - Only 2 questions resulted in different queries

<table>
<thead>
<tr>
<th></th>
<th>R-Prec</th>
<th>MAP</th>
<th>Exact Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Query</td>
<td>0.4464</td>
<td>0.4357</td>
<td>0.1443</td>
</tr>
<tr>
<td>Query w/ Disambiguation</td>
<td>0.4658</td>
<td>0.4409</td>
<td>0.1443</td>
</tr>
</tbody>
</table>
Search Term Context Specification

- **Experimental setup:**
  - Test set: 46 questions from the AQUAINT 2005 opinion pilot of the general form “What does A think about B?”
  - Baseline query:
    - OpinionHolder + Opinion
  - Query with context:
    - OpinionHolder + <Quotation> Opinion </Quotation>

- **Evaluation Results**

<table>
<thead>
<tr>
<th></th>
<th># Vital Nuggets</th>
<th># Okay Nuggets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Query</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>Query w/ Context</td>
<td>28</td>
<td>5</td>
</tr>
</tbody>
</table>
Relation Specification Evaluation

- **Experimental setup:**
  - Test set: 25 semantic queries constructed from 10 relations in the national intelligence domain
  - Relation query: constrained semantic search query with relation
    - `<ProducesWeapon>+Russia +<ChemicalWeapon></ChemicalWeapon>`
  - Baseline query: relation substituted by keywords
    - `+Russia +<ChemicalWeapon></ChemicalWeapon>` produce

- **Evaluation Results**

<table>
<thead>
<tr>
<th></th>
<th>R-Prec</th>
<th>MAP</th>
<th>Exact Precision</th>
<th># docs/Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Query</td>
<td>0.3895</td>
<td>0.4147</td>
<td>0.3530</td>
<td>9.72</td>
</tr>
<tr>
<td>Query w/ Relation</td>
<td>0.4139</td>
<td>0.4108</td>
<td>0.6108</td>
<td>6.32</td>
</tr>
</tbody>
</table>
SAW: Semantic Analysis Workbench
- User can type keywords, phrases, or questions
- Query Analysis generates Semantic Search Query
- Intelligent passage selection
- Hones in on relevant sections
- Redefines the right grain
• Keyword matches only
• Patch-work passages
Eye on Proliferation: WMD Country Profiles

Iraq Profile

Search Iraq:

Facilities  Chronology  Import/Export

Nuclear Overview

Iraq came very close to becoming the first Arab country to produce nuclear weapons. Before its invasion of Kuwait, Iraq, a signatory to the Nuclear Non-Proliferation Treaty, had no nuclear weapons. In 1991, it may have been able to build a weapon within a few years. With a bomb design progressing rapidly, Iraq lacked only the fissile material.

The Israeli "Threat"

Iraq maintained a very small civilian nuclear program that began during the Atomic for Peace program in the mid-1950s. In 1962, it acquired a 2 megawatt (MW) research reactor. In 1974, an Iraqi delegation traveled to Paris to negotiate the purchase of a reactor similar to the French Cauris reactor. Then Vice-President Saddam Hussein would have preferred to purchase the U.S. PFI reactor, but had to settle for the Cauris reactor instead. While Iraq attempted to maintain the facade that its purchase was for peaceful nuclear research, some of its other actions belied stated intentions. Prior to his trip, Saddam sought to purchase a reactor from the United States.

With the reactor deal complete, Iraq set out to build a radiochemical laboratory. It contracted the Italian firm SNIA-Techint in 1979 to build a pilot plutonium reactor. This period also saw the return to Iran of Ayatollah Ruholla Khomeini, the Shiite cleric, from his exile in Iraq. Shortly thereafter, Iran became an Islamic republic.

The 10-year war with Iran played a role in Saddam's thinking vis-a-vis Iraq's nuclear weapon program. Saddam wanted the "ultimate equalizer" if faced with an overwhelming foe. Iraqi scientists calculated that the Osirak reactor could produce between 5 and 25 kilograms (kg) of plutonium per year (the upper limit, however, would require the use of enriched uranium).

Iraqi scientists calculated that the Osirak reactor could produce between 5 and 25 kilograms (kg) of plutonium per year (the upper limit, however, would require the use of enriched uranium).

Iraq proceeded with building its nuclear infrastructure and plans to subvert IAEA inspections. However, before any significant progress could be made, the Osirak reactor was destroyed before it could produce any plutonium. Iraq had to seek another source of fissile material. Uranium enrichment became the priority. As a secondary method, Iraq explored uranium enrichment through the gaseous diffusion process. The intention was to produce low enriched uranium as feedstock for use in a reactor.
• Precision query exploits relations
• 49 results (rather than 500+) using highly relevant relation
- Annotated document view
- Reveals task-relevant concepts and relations
- Tracks source of annotation
• User want to get more specific
• Based on task ontology gets options for “restricting” selected concepts
• Goes from Weapon to Chemical Weapon
Zoomed in on single highly relevant document
The image shows a software interface for natural language processing or information extraction. The main window displays a query for identifying weapons produced by Iraq. There are annotations pointing to specific phrases and entities:

- Germany
- ProducesWeapon
- ChemicalWeapon
- PoisonGasPlants
- Iraq

The software appears to be analyzing text to extract information about weapons production, specifically focusing on the chemical and poison gas plants in Iraq.
• Attempts “restriction” with Google
• No reduction/precision gain
• Finds documents with “Chemical Weapon”
Conclusions

- **Semantic Search enables high precision IR**
  - Identified three XML Fragments operators
  - Applied operators to address four query time semantic needs
  - Shown significant improvement in precision in most applications

- **Semantic Search not intended as replacement for keyword search**
  - Keyword search is successful in web and desktop search
  - Semantic search intended to improve performance for
    - Enterprise search
    - Information gathering for exploratory research