Question Answering System for Yioop

Advisor
Dr. Chris Pollett

Committee Members
Dr. Thomas Austin
Dr. Robert Chun

By
Niravkumar Patel
Outline

- Problem Statement
- Question Answering System
- Yioop
- Proposed System
- Triplet Extraction Approach
- Integration with Yioop
- Observations
- Areas of Improvement
- Conclusion
Problem Statement

• Information Retrieval in Yioop doesn’t provide a specific answer to the query entered in the form of a question
• Currently, Yioop treats both a query statement and a question in the same manner
Example 1
Example 2
Are questions really asked by users online?

- Search engines query log analysis shows that

<table>
<thead>
<tr>
<th>Types of Query</th>
<th>Query log analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informational</td>
<td>48%</td>
</tr>
<tr>
<td>Navigational</td>
<td>20%</td>
</tr>
<tr>
<td>Transactional</td>
<td>30%</td>
</tr>
</tbody>
</table>

- 27% of Informational queries are questions
Question Answering System

### START

**Natural Language Question Answering System**

`where is paris`

The coordinates of Paris, France are 48.86 N, 2.33 E.

**Source:** START KB

Paris, France is located at 114.0 feet above sea level.

**Source:** Global Gazetteer

---

**Who is Christopher Columbus**

Christopher Columbus was an Italian explorer, navigator, colonizer and citizen of the Republic of Genoa. Under the auspices of the Catholic Monarchs of Spain, he completed four voyages across the Atlantic Ocean. Those voyages led directly to the discovery of the New World and the beginning of the Columbian Exchange. Columbus' last voyage, the 1493-1496 voyage, was sponsored by the Spanish crown to find a new route to Asia.

**Source:** Wikipedia, the free encyclopedia

---

**Who is the President of the United States**

Barack Obama

United States of America / President

Barack Obama became the 44th and current president on November 6, 2012. He was re-elected and is currently serving the 57th term, which ends in January 20, 2021.
Yioop

• Open Source Search Engine
• Major processes
  – Crawler
  – Summarizer
  – Index builder
  – Query Parser
Information Flow in Yioop

Query Parser → Inverted Index

Fetch data from documents

Summarizer
Indexer

World Wide Web
Proposed Question Answering System for Yioop

Query Parser

Inverted Index

Question Answering System

Summarizer

Indexer

World Wide Web

Fetch data from documents
Approaches for Question Answering System

- Named-entity recognition technology
- Knowledge based Approach
- Triplet Extraction Approach
Approach Chosen

• Triplet Extraction Algorithm
• Identifies extraction of relations between Subject and Object
• [SUBJECT – PREDICATE - OBJECT]
Prerequisite

• Structure of statement
• Part Of Speech Tagger
• Parse Tree Generation
Structure of Statement

• The phrase given to the system should be a complete statement.

• Ex.
  – “Barack Obama was born on August 4 1961” => [Barack Obama – born – on August 4 1961]
Part of Speech Tagger

• Example
  – Statement: The grand jury commented on a number of other topics

• Output of statement

• Tagging Technique used
  – Brill Tagger
Parse Tree Generation

• Grammar rules to interpret a common form of a statement

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NP:</td>
<td>{&lt;DT</td>
</tr>
<tr>
<td>PP:</td>
<td>{&lt;IN&gt;&lt;NP&gt;}</td>
</tr>
<tr>
<td>VP:</td>
<td>{&lt;VB.*&gt;&lt;NP</td>
</tr>
<tr>
<td>Statement:</td>
<td>{&lt;NP&gt;&lt;VP&gt;}</td>
</tr>
</tbody>
</table>
Triplet Extraction Algorithm

- **Subject Extraction**: Noun in the NP Sub-tree
- **Predicate Extraction**: Deepest Verb descendent of Verb Phrase
- **Object Extraction**: All siblings of VB in Verb Phrase Sub-tree
Triplet Extraction Algorithm

• Statement: Alice chased the rabbit
• POS output: Alice [NN] chased [VB] the [DT] rabbit [NN]
• Triplet Extracted: [ALICE – CHASED – THE RABBIT]
Ex. Alice chased the rabbit

```
Statement

NP
NN
Alice

VP
VB
Chased

NP
DT
The
NN
Rabbit
```
Ex. Mary saw the black cat

```
NP
  |  
  |  
  |  
NN  
  |  
  |  
VP
  |  
  |  
  |  
NP
  |  
  |  
  |  
DT  JJ  NN
  |  |  
  |  |  
  |  |  
The  Black  Cat
```
Ex. Narendra Modi is the Prime Minister of India
Integration with Yioop

• At Crawl time
• At Query time
Integration at Crawl time

After page rules applied

Words and positions extracted to index from summary

Hypertext Transfer Protocol

From Wikipedia, the free encyclopedia

The Hypertext Transfer Protocol (HTTP) is an application protocol for distributed, collaborative, hypermedia information systems. HTTP is the foundation of data communication for the World Wide Web.

HTTP is a stateless protocol that uses logical links (hyperlinks) between nodes containing text. HTTP is the protocol to exchange or transfer hypertext.

The standards development of HTTP was coordinated by the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C), culminating in the publication of a series of Requests for Comments (RFCs). The first definitive version of HTTP/1.1, the version of HTTP in common use, was documented in RFC 2616 in 1999, although this was obsolesced by RFC 7230 in 2014.
[http is qque] ⇒ a stateless protocol
[http qque a Stateless protocol] ⇒ as http
[http qque a stateless protocol] ⇒ is
[qque be us] ⇒ challeng respons authent scheme
[authent scheme qque us] ⇒ be
[authent scheme be qque] ⇒ us by a server
[qque be us by a server] ⇒ scheme
[authent scheme qque us by a server] ⇒ be
[qque digest access] ⇒ authent scheme such
[authent scheme qque access] ⇒ digest
[authent scheme digest qque] ⇒ access
[authent scheme such qque access] ⇒ digest
[authent scheme such digest qque] ⇒ access
[qque digest access authent] ⇒ access authent
[authent qque access authent] ⇒ digest
[authent digest qque] ⇒ access authent
[qque set of challeng respons authent scheme] ⇒ extens
Limitation on Triplet Extraction

• Number of words
• Fails to ignore triplets having unnecessary Subject/Object
  – [Nicholas Carsen – Wrote – In 2014]
  – [The university website – Made - Available]
Integration at Query Time

- **Ex.**
  - What is http? => HTTP is *qqque*
  - Who ate the dog? => *qqque* ate the dog
  - Where is SJSU University located? => [SJSU University - is located – *qqque*] & [SJSU University - located – *qqque*]

(*qqque* is the identifier for the question word)
Limitation on Question Parser

- Processes question that starts with Who, Where, What.
- Ex.
  - Why is the sky blue?
  - How many days in a year?
  - President of USA is?
Observation on results

Query results on Yioop without Question Answering System

Query results on Yioop with Question Answering System
Observation on performance

• Trade off between space and time
  – [Subject – Predicate – Object] -> “Offset information in Paragraph”
  – [Subject – Predicate – Object] -> “Answer”
Areas of Improvement

• Overall accuracy of Question Answering System depends on the accuracy of individual components
• Junk triplet removal strategy
• User feedback
• Limitation on the Source of Information for Question Answering System
Conclusion

• Question Answering System will help Yioop in answering the questions asked by the user
• Performance lag only at crawl time
• Integration with Web Interface helps the user/developer to see the triplets generated for individual web pages
Questions?