A SEMANTIC KNOWLEDGE ENGINE USING AUTOMATED KNOWLEDGE EXTRACTION FROM WORLD WIDE WEB

A Project by

Venkatesh Mabbu

Bachelor of Engineering and Technology, Jawaharlal Nehru Technological University, 2013

Submitted to the Department of Electrical Engineering and Computer Science
and the faculty of the Graduate School of
Wichita State University
in partial fulfillment of
the requirements for the degree of
Master of Science

December 2015

A SEMANTIC KNOWLEDGE ENGINE USING AUTOMATED KNOWLEDGE

EXTRACTION FROM WORLD WIDE WEB

The following faculty members have examined the final copy of this report for form and content, and recommends that it be accepted in partial fulfillment of the requirement for the degree of Master of Science with a major in Computer Science.

Dr. Abu Asaduzzaman, Committee Chair

Dr. Kaushik Sinha, Committee Member

Dr. Ramazan Asmatulu, Committee Member

DEDICATION

TO MY PARENTS AND GRANDPARENTS

ACKNOWLEDGEMENTS

I am thankful to my project advisor Dr. Abu Asaduzzaman for his encouragement, support, and supervision throughout this work. His timely guidance to my work helped this research to be completed on time. He always had patience to advise me in accomplishing this work in spite of his busy schedule and offered me assistance from time to time. I also express my gratitude and thanks towards Dr. Kaushik Sinha (WSU Electrical Engineering and Computer Science Department) and Dr. Ramazan Asmatulu (WSU Mechanical Engineering Department) for taking time from their busy schedules and serve in my MS Project committee.

I take pleasure in recognizing the efforts of all those who encouraged and assisted me both directly and indirectly with my work and experiments. I acknowledge WSU Computer Architecture and Parallel Program Laboratory (CAPPLab) research group for providing valuable suggestions to improve the quality of the project and this manuscript

ABSTRACT

It becomes extremely difficult for the existing search engines (such as Google, Bing, and Yandex) to crawl, index, rank, and manage huge amount of data and locate information while answering questions. Semantic web technology (such as Google Knowledge Graph, WolframAlpha, Freebase, and Wikidata) is emerging into the answer engine market in order to transform the unstructured data into more structured useful information. However, the existing engines suffer due to the fact that curators and volunteers feed these systems manually. In this project, we aim to transform the unstructured data into more useful data using an automation technique. We implement the proposed system in 20 different categories including universities. Based on a survey among 50 university students, we receive excellent satisfactory results as the proposed engine answers more effectively. In an average, the proposed engine energy consumption, search time and storage is 1 million times lesser than the existing search engines (see Section 5.1).

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

LIST OF ABBREVIATIONS

CAPPLab	Computer Architecture and Parallel Programming Laboratory
DMCA	Digital Millennium Copyright Act
HTML	Hypertext Markup Language
JSON	Javascript Object Notation
RDF	Resource Description Framework
SEO	Search Engine Optimization
SERP	Search Engine Results Page
URL	Uniform Resource Locator
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
XML	Extensible Markup Language

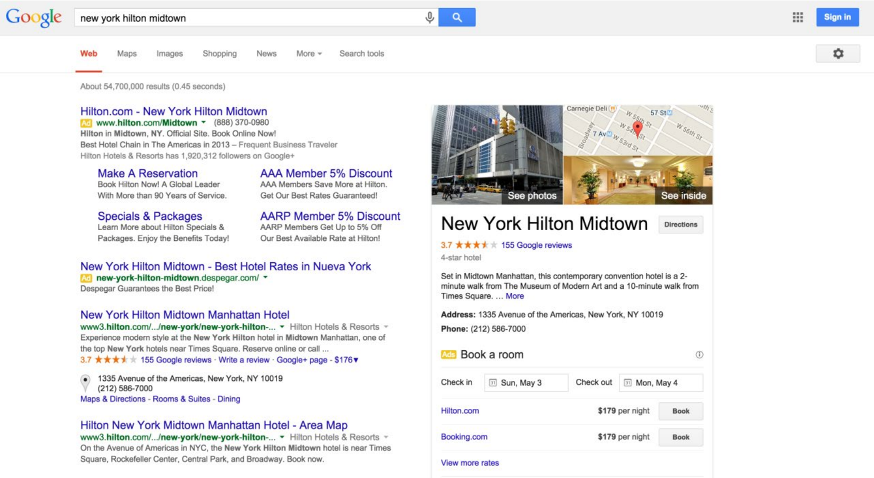
CHAPTER 1

INTRODUCTION

This chapter gives an overview about various search engines [1] like Google [2] and semantic engines like Wolfram Alpha [3] and structured web projects like Freebase, Wikidata and DBpedia.

1.1 Search Engine

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.



About 54,700,000 results (0.45 seconds)

Hilton.com - New York Hilton Midtown
Ad www.hilton.com/Midtown (888) 370-0980
Hilton in Midtown, NY. Official Site. Book Online Now!
Best Hotel Chain in The Americas in 2013 – Frequent Business Traveler
Hilton Hotels & Resorts has 1,920,312 followers on Google+

Make A Reservation
Book Hilton Now! A Global Leader With More than 90 Years of Service.

AAA Member 5% Discount
AAA Members Save More at Hilton. Get Our Best Rates Guaranteed!

Specials & Packages
Learn More about Hilton Specials & Packages. Enjoy the Benefits Today!

AARP Member 5% Discount
AARP Members Get Up to 5% Off Our Best Available Rate at Hilton!

New York Hilton Midtown - Best Hotel Rates in Nueva York
Ad new-york-hilton-midtown.despegar.com/
Despegar Guarantees the Best Price!

New York Hilton Midtown Manhattan Hotel
www3.hilton.com/.../new-york/new-york-hilton-... Hilton Hotels & Resorts
Experience modern style at the New York Hilton hotel in Midtown Manhattan, one of the top New York hotels near Times Square. Reserve online or call ...
3.7 ★★★★☆ 155 Google reviews · Write a review · Google+ page - $176
1335 Avenue of the Americas, New York, NY 10019
(212) 586-7000
Maps & Directions - Rooms & Suites - Dining

Hilton New York Midtown Manhattan Hotel - Area Map
www3.hilton.com/.../new-york/new-york-hilton-... Hilton Hotels & Resorts
On the Avenue of Americas in NYC, the New York Hilton Midtown hotel is near Times Square, Rockefeller Center, Central Park, and Broadway. Book now.

New York Hilton Midtown
3.7 ★★★★☆ 155 Google reviews
4-star hotel
Set in Midtown Manhattan, this contemporary convention hotel is a 2-minute walk from The Museum of Modern Art and a 10-minute walk from Times Square. ... More
Address: 1335 Avenue of the Americas, New York, NY 10019
Phone: (212) 586-7000

Ads Book a room
Check in Sun, May 3 Check out Mon, May 4
Hilton.com $179 per night
Booking.com $179 per night
View more rates

Fig. 1.1: Screenshot of Google Search Engine Results Page

1.2 Semantic Search

Semantic search [8][9] seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data space, whether on the Web or within a closed system, to generate more relevant results.

The following figure shows an overview of semantic search, every keyword treated as thing instead of a string and all these nodes are interconnected in a semantic way such that the semantic search capable of answering the questions by just looking at attributes instead of the documents containing the keywords the user is looking for.

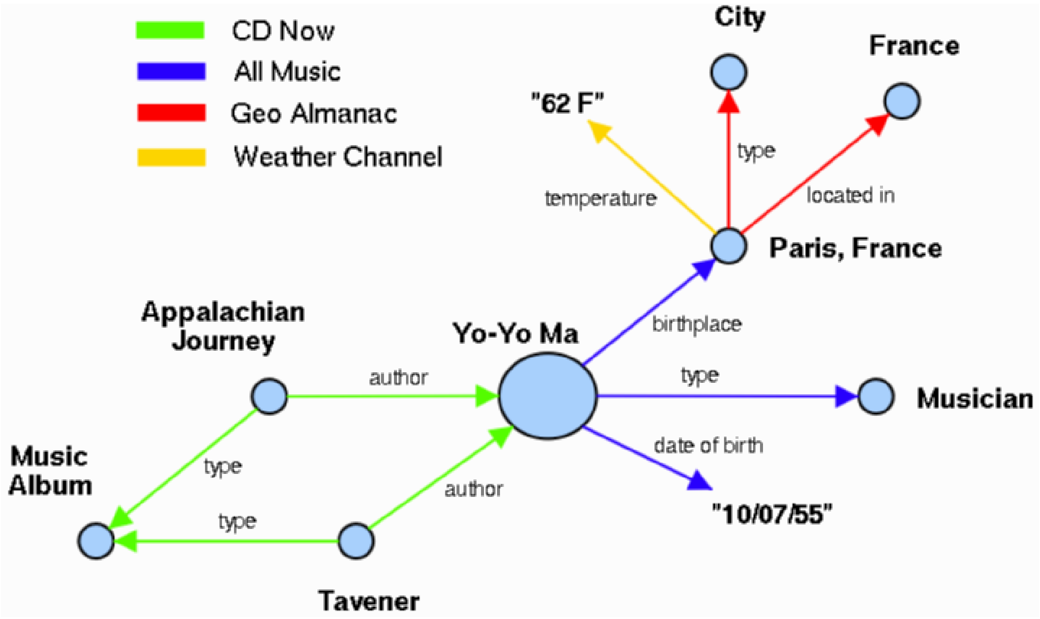
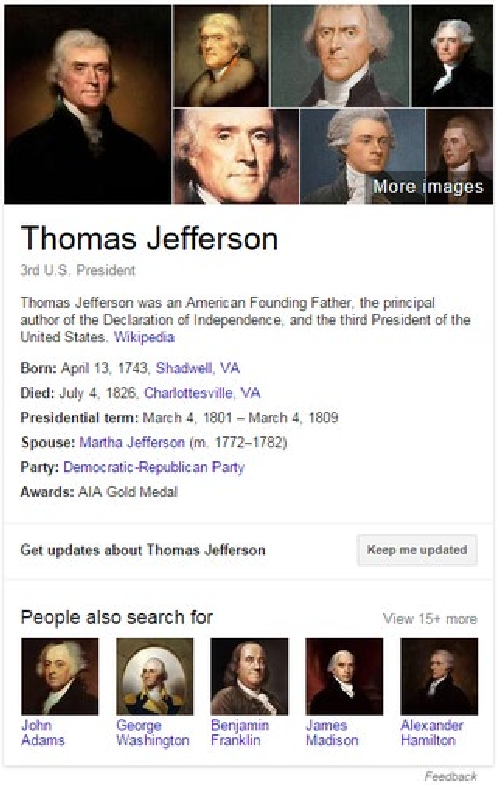


Fig. 1.2: Overview of Semantic Web

1.3 Knowledge Graph

The Knowledge Graph [5] is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources. Knowledge Graph display was added to Google's search engine in 2012, starting in the United States. According to Google, the information in the Knowledge Graph is derived from many sources, including the CIA World Factbook, Freebase [6] (being replaced by Wikidata[7]), and Wikipedia. The feature is similar in intent to answer engines such as Wolfram Alpha and efforts such as Linked Data and DBpedia.



Thomas Jefferson

3rd U.S. President

Thomas Jefferson was an American Founding Father, the principal author of the Declaration of Independence, and the third President of the United States. Wikipedia

Born: April 13, 1743, Shadwell, VA

Died: July 4, 1826, Charlottesville, VA

Presidential term: March 4, 1801 – March 4, 1809

Spouse: Martha Jefferson (m. 1772–1782)

Party: Democratic-Republican Party

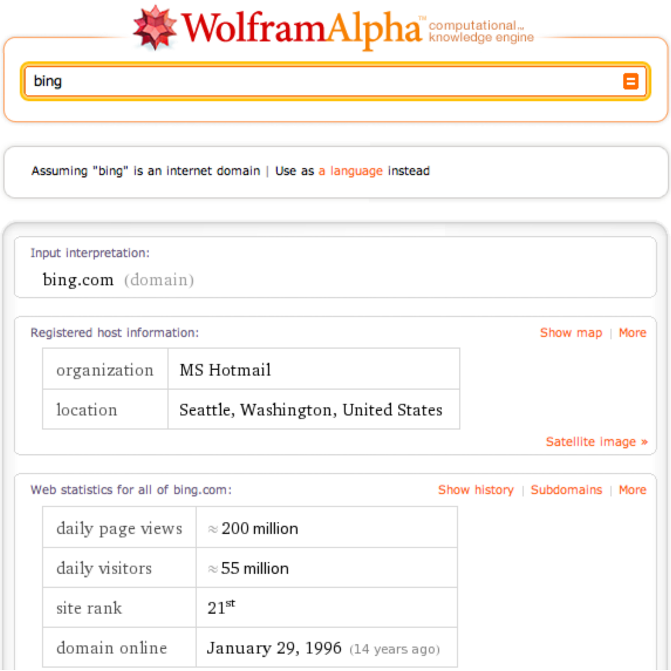
Awards: AIA Gold Medal

Get updates about Thomas Jefferson

Fig. 1.3: Knowledge Graph data about Thomas Jefferson displayed on Google Web Search

1.4 Wolfram Alpha

Wolfram Alpha is a computational knowledge engine or answer engine developed by Wolfram Research. It is an online service that answers factual queries directly by computing the answer from externally sourced "curated data", rather than providing a list of documents or web pages that might contain the answer as a search engine might. The curated data makes Alpha different from semantic search engines [29][31], which index a large number of answers and then try to match the question to one. The following diagram shows the result on Wolfram Alpha search engine for the keyword “bing.”


WolframAlpha™ computational.. knowledge engine
bing
Show map | More
Satellite image »
Show history | Subdomains | More

Assuming "bing" is an internet domain | Use as a language instead

Input interpretation:
bing.com (domain)

Registered host information:

organization	MS Hotmail
location	Seattle, Washington, United States

Web statistics for all of bing.com:

daily page views	≈ 200 million
daily visitors	≈ 55 million
site rank	21st
domain online	January 29, 1996 (14 years ago)

Fig. 1.4: Wolfram Alpha Results Page

1.5 DBpedia

DBpedia [9][10][14] (from "DB" for "database") is a project aiming to extract structured content from the information created as part of the Wikipedia project. This structured information is then made available on the World Wide Web. DBpedia allows users to semantically query relationships and properties associated with Wikipedia resources, including links to other related datasets.

1.6 Freebase

Freebase was a large collaborative knowledge base consisting of data composed mainly by its community members. It was an online collection of structured data harvested from many sources, including individual, user-submitted wiki contributions. Freebase aimed to create a global resource that allowed people (and machines) to access common information more effectively.

1.7 Wikidata

Wikidata is a collaboratively edited knowledge base operated by the Wikimedia Foundation. It is intended to provide a common source of certain data types (for example, birth dates) which can be used by Wikimedia projects such as Wikipedia. This is similar to the way Wikimedia Commons provides storage for media files and access to those files for all Wikimedia projects. Wikidata is powered by the software Wikibase.

1.8 TrueKnowledge

Evi (formerly True Knowledge) is a technology company in Cambridge, England, founded by William Tunstall-Pedoe, which specialises in knowledge base and semantic search engine software. Its first product was an answer engine that aimed to directly answer questions posed in plain English text, which is accomplished using a database of discrete facts [15]. The True Knowledge Answer engine was launched for private beta testing and development on 7 November 2007.

CHAPTER 2

PROBLEM DESCRIPTION AND CONTRIBUTIONS

This chapter discusses about the current issues in the field of “Search Engine” motivated for this research.

2.1 Problem Description

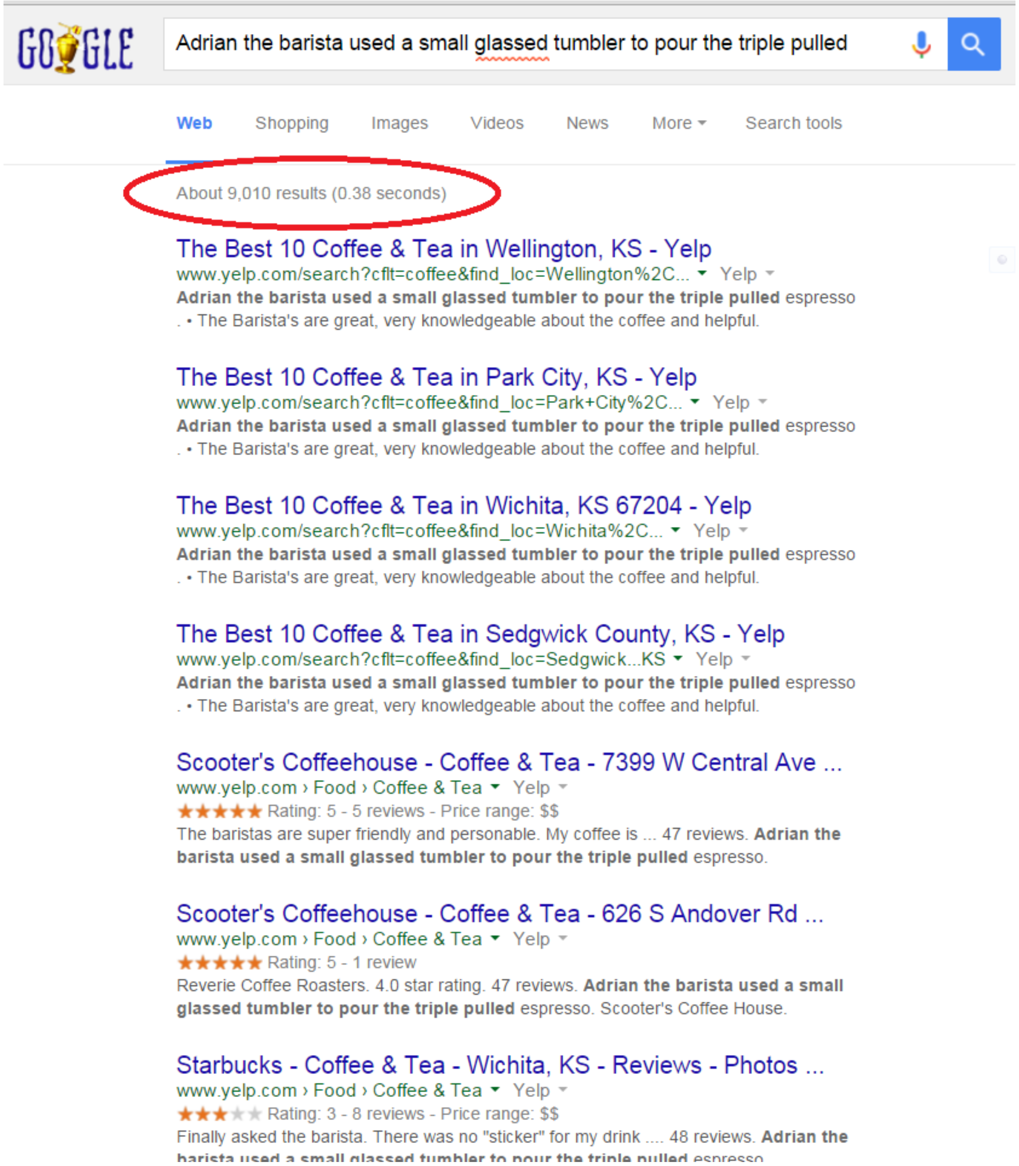
It is often a very big problem for search engines like Google to crawl, index, summarize and monitor the vast World Wide Web. Based on the Google indexed data, it is found that there are 50 billion unique web pages on World Wide Web. In most of these web pages, the information is highly redundant as so many web pages discussing about exactly the same content in a different URL. For search engines like Google, it is difficult to classify and cluster the URL which are discussing the same information. Among these vast useful web pages, there are many spammed/illegal web pages which may harm the user. The search engines are trying hard in eliminating these web pages from their search results based on many Machine Learning techniques.

In order to provide most accurate results, search engines always tries to track many private information from the users like their location, search/browsing habits, websites they visited and time he/she spent on each website which leads to raise many concerns from the users about their privacy. Another problem is “Copyright infringement”, some websites on World Wide Web hosts the illegal/copyright data available to the users for download, this piracy brings enormous loss to so many industries like Software companies, Film companies, etc. Google used to receive highest number of complaints from DMCA every year to remove the illegal links from its search results page.

The figure 2.1 is a screenshot of Google Search Result showing 9010 results for one single review written on "yelp." This is a small example which demonstrates how much redundancy exists on World Wide Web. Yelp provides have their own search services but search engines still index those dynamic websites which contains the same information but on different URL and thereby wasting valuable resources.

There are many reasons for redundant data on World Wide Web some are because of dynamic nature of the websites like Yelp, Quora, Airbnb, etc. Which produces information on demand (i.e., based on the user query) Being a crawler like Google access the website in every possible pages which may be a different URL name but with same content.

Another reason is, designing and publishing a website or blog became so simple these days, so majority of the web masters are starting their own websites and blogs with the data which is just a grouping of data from different websites. Because of these websites the size of internet is becoming enormously.


Adrian the barista used a small glassed tumbler to pour the triple pulled
Web
Shopping
Images
Videos
News
More
Search tools

About 9,010 results (0.38 seconds)

The Best 10 Coffee & Tea in Wellington, KS - Yelp
www.yelp.com/search?cflt=coffee&find_loc=Wellington%2C... Yelp
Adrian the barista used a small glassed tumbler to pour the triple pulled espresso . • The Barista's are great, very knowledgeable about the coffee and helpful.

The Best 10 Coffee & Tea in Park City, KS - Yelp
www.yelp.com/search?cflt=coffee&find_loc=Park+City%2C... Yelp
Adrian the barista used a small glassed tumbler to pour the triple pulled espresso . • The Barista's are great, very knowledgeable about the coffee and helpful.

The Best 10 Coffee & Tea in Wichita, KS 67204 - Yelp
www.yelp.com/search?cflt=coffee&find_loc=Wichita%2C... Yelp
Adrian the barista used a small glassed tumbler to pour the triple pulled espresso . • The Barista's are great, very knowledgeable about the coffee and helpful.

The Best 10 Coffee & Tea in Sedgwick County, KS - Yelp
www.yelp.com/search?cflt=coffee&find_loc=Sedgwick...KS Yelp
Adrian the barista used a small glassed tumbler to pour the triple pulled espresso . • The Barista's are great, very knowledgeable about the coffee and helpful.

Scooter's Coffeehouse - Coffee & Tea - 7399 W Central Ave ...
www.yelp.com › Food › Coffee & Tea Yelp
★★★★★ Rating: 5 - 5 reviews - Price range: $$
The baristas are super friendly and personable. My coffee is ... 47 reviews. **Adrian the barista used a small glassed tumbler to pour the triple pulled** espresso.

Scooter's Coffeehouse - Coffee & Tea - 626 S Andover Rd ...
www.yelp.com › Food › Coffee & Tea Yelp
★★★★★ Rating: 5 - 1 review
Reverie Coffee Roasters. 4.0 star rating. 47 reviews. **Adrian the barista used a small glassed tumbler to pour the triple pulled** espresso. Scooter's Coffee House.

Starbucks - Coffee & Tea - Wichita, KS - Reviews - Photos ...
www.yelp.com › Food › Coffee & Tea Yelp
★★★☆☆ Rating: 3 - 8 reviews - Price range: $$
Finally asked the barista. There was no "sticker" for my drink .... 48 reviews. **Adrian the barista used a small glassed tumbler to pour the triple pulled** espresso

Fig. 2.1: Google SERP for a Yelp review

For the query “How to tie a tie”, Google returned 102,000,000 results just discussing about a procedure of “how to tie a tie." One of the first step like "Start with the wide end of the tie on the right and the small end on the left" produced "132,000,000", this example helps in understanding how much redundant information exist on world wide web.

For another query "houses for rent in wichita ks", Google returned 1,310,000 results but population of Wichita is just 400k. Every property in the list is duplicated at least 3000 times in different websites. If this trend continues we may witness even a million duplicates for just one single house in the future. We are wasting our time, energy and human personnel to crawl, index and manage this redundant information on World Wide Web.

The simple solution for this problem is "Semantic Search." In this methodology the system understands all the possible information about the real-world and the programs are well trained to collect only the specific information from more specific sources. This will help the system capable of answering user queries in the best efficient manner when compared with any other search engines by just utilizing 1 millionth part of resources. The semantic search is at very early stage and the companies are implementing their defining their own strategies in order to provide the best answer to the user.

2.2 Project Contributions

This project is focusing on all the pre-discussed issues and proposing the best possible solution in overcoming these problems. The following are the main contribution of this project:

a. **Reduces the size of index:** By using this Semantic Search we can able to reduce the index size to one millionth times when compared with regular search engines without compromising any quality.

b. **Reduces the redundant data:** We noticed that one fact in World Wide Web is being duplicated for thousands of times and by using this semantic search we will index "one fact one time" which may completely prevent the redundant results from the search results.

c. Enhance Privacy: The Semantic Search completely works on the context of user search and the system is enough knowledgeable to answer the user queries without any external factors like user browsing/search habits, IP, etc. So Semantic Search doesn't need to track user activities in any form which may help users stop worrying about their Privacy.

d. Protects Copyrights: The most knowledgeable engine like Semantic Search engines just direct the users towards the legitimate sites for downloading/purchasing a film/software and it doesn't crawls and manage the data of illegal websites. So, the Semantic Search users never get an illegal website link in their results thereby protects the copyrights of the respective owners.

e. **No access to Spamming Websites:** As discussed earlier, the Semantic Search only crawls and indexes the selective most genuine websites from the information so user wil never et a link to the spamming website which may harm them.

CHAPTER 3

LITERATURE SURVEY

This chapter elaborates previous and related work that had been in the research till date. Various technologies that have been using in the areas of

3.1 Resource Description Framework

RDF [9] is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. The RDF data model is similar to classical conceptual modeling approaches such as entity–relationship or class diagrams, as it is based upon the idea of making statements about resources (in particular web resources) in the form of subject–predicate–object expressions. These expressions are known as triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

A RDF tuple has following characteristics:

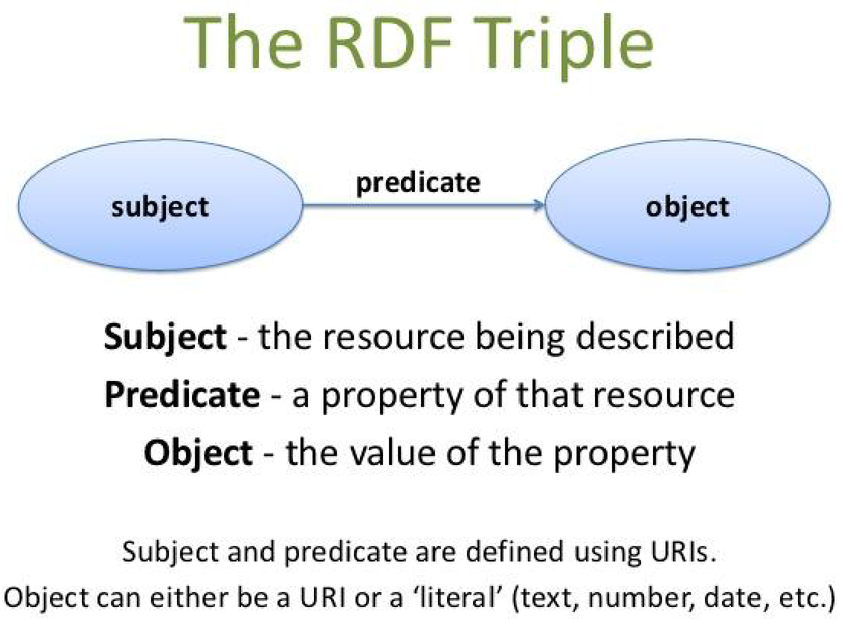
a. Each RDF triple is made up of subject, predicate and object.

b. Each RDF triple is a complete and unique fact.

c. An RDF triple is a 3-tuple, which is made up of a subject, predicate and object – which are respectively an uriref or bnode; an uriref; and an uriref, bnode or literal.

d. Each RDF triple can be joined with other RDF triples, but it still retains its own unique meaning, regardless of the complexity of the models in which it is included.

The figure 3.1 is an example of RDF triple with subject, object and predicate. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.



The RDF Triple

Subject - the resource being described

Predicate - a property of that resource

Object - the value of the property

Subject and predicate are defined using URIs.
Object can either be a URI or a 'literal' (text, number, date, etc.)

Fig. 3.1: RDF Triple

Sample RDF/XML Fromat:

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#"
xmlns:eric="http://www.w3.org/People/EM/contact#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:mailbox rdf:resource="mailto:e.miller123(at)example"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:personalTitle>Dr.</contact:personalTitle>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <rdf:type
rdf:resource="http://www.w3.org/2000/10/swap/pim/contact#Person"/>
  </rdf:Description>
</rdf:RDF>
```

Fig. 3.2: Sample RDF/XML Format

The above diagram is describing a resource with statements there is a Person identified by http://www.w3.org/People/EM/contact#me, whose name is Eric Miller, whose email address is e.miller123(at)example (changed for security purposes), and whose title is Dr.

Subject: The resource "http://www.w3.org/People/EM/contact#me" is the subject.

Objects are: "Eric Miller" (name),

Predicate: "whose name is" is a predicate.

The subject is a URI. The predicates also have URIs. For example, the URI for each predicate:

"whose name is" is “http://www.w3.org/2000/10/swap/pim/contact#fullName”

3.1.1 Serialization formats

Several common serialization formats are in use, including:

Turtle: a compact, human-friendly format.

N-Triples: a very simple, easy-to-parse, line-based format that is not as compact as Turtle.

N-Quads: a superset of N-Triples, for serializing multiple RDF graphs.

JSON-LD: a JSON-based serialization.

3.2 Schemas

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet[17], on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

3.2.1 Examples of schema

The following is an example of how to mark up information about a movie and its director using the schema.org schemas and microdata. In order to mark up the data the attribute itemtype along with the URL of the schema is used. The attribute itemscope defines the scope of the itemtype. The kind of the current item can be defined by using the attribute itemprop.

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
  Director: <span itemprop="name">James Cameron</span>
(born <time itemprop="birthDate" datetime="1954-08-16">August 16,
1954</time>)
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html"
itemprop="trailer">Trailer</a>
</div>
```

Fig. 3.3: Example of Microdata Schema Representation

```
<div vocab="http://schema.org/" typeof="Movie">
  <h1 property="name">Avatar</h1>
  <div property="director" typeof="Person">
  Director: <span property="name">James Cameron</span>
```

```
(born <time property="birthDate" datetime="1954-08-16">August 16,
1954</time>)
  </div>
  <span property="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html"
property="trailer">Trailer</a>
</div>
```

Fig. 3.4: Example of RDFa 1.1 Lite Schema Representation

```
<script type="application/ld+json">
{
  "@context": "http://schema.org/",
  "@type": "Movie",
  "name": "Avatar",
  "director":
    {
        "@type": "Person",
        "name": "James Cameron",
        "birthDate": "1954-08-16"
    },
  "genre": "Science fiction",
  "trailer": "../movies/avatar-theatrical-trailer.html"
}
</script>
```

Fig. 3.5: Example of JSON-LD Schema Representation

3.3 Limitations

3.3.1 Schema types are limited.

Structured data is great for people, products, places, and events, but these cover only a fraction of the entire content of the web. Many of us markup our content using Article schema, but this falls well short of describing the hundreds of possible entity associations within the text itself.

3.3.2 Markup is difficult

Realistically, in a world where it's sometimes difficult to get authors to write a title tag or get engineers to attach an alt attribute to an image, implementing proper structured data to source HTML can be a daunting task [12].

3.3.3 Adoption is low

A study last year of 2.4 billion web pages showed less than 25% contained structured data markup [18]. A recent Search Metrics study showed even less adoption, with only 0.3% of websites out of over 50 million domains using Schema.org.

CHAPTER 4

PROPOSED SYSTEM

The Semantic Web is an extension of the current Web that allows the meaning of information to be precisely described in terms of well-defined vocabularies that are understood by people and computers. On the Semantic Web information is described using a new W3C standard called the Resource Description Framework (RDF). Currently research on semantic web search engines are in the beginning stage, as the traditional search engines such as Google, Yahoo, and Bing (MSN) and so forth still dominate the present markets of search engines. Current web is the biggest global database that lacks the existence of a semantic structure and hence it makes difficult for the machine to understand the information provided by the user.

When the information was distributed in web, we have two kinds of research problems in search engine i.e.

a. How can a search engine map a query to documents where information is available but does not retrieve in intelligent and meaningful information?
b. The query results produced by search engines are distributed across different documents that may be connected with hyperlink. How search engine can recognize efficiently such a distributed results?

Semantic web can solve the first problem in web with semantic annotations to produce intelligent and meaningful information by using query interface mechanism and ontologies. Other one can be solved by the graph-based query models. The Semantic web would require solving extraordinarily difficult problems in the areas of knowledge representation, natural language understanding. Currently many of semantic search engines are developed and implemented in different working environments, and these mechanisms can be put into use to realize present search engines.

The proposed system not just concentrates on Semantic Search, it actually a combination of both Semantic and Regular Web Search designed to answer majority of user queries. For this, we designed a prototype based on user search habits. We identified three popular different searches a user performs on internet.

a. **Instant Real Time Data:** The queries which come under this category are definitions, weather, Stock market data, different conversions (currency, area, volume, temperature, etc.), basic calculations (like mathematical expression solver)
b. **Factual Data:** This project is highly focusing on this section as it requires intelligent and smart computations in order to answer the user query. For this we indexed early one million facts in eleven different categories from the most reliable sources on internet like Wikipedia, CIA world Factbook, us news, US Department of Agriculture and more popular websites.
c. **Web Results:** This section of queries needs to be answered with hyperlinks instead of answers like Cheap flights, amazon, manage Bank of America account, etc.

In order to answer the user questions semantically, first the system should be trained with all the structured information available on web. As the web is unstructured, we designed a program which automatically transforms this unstructured data into most useful and easily accessible knowledge using **smart automated bot**. The proposed system is world's first of its kind in transforming unstructured data into structured web with ***zero human intervention***. The system is trained with highly structured millions of entities and attributes in order to produce a series of facts and those facts will help in solving user posed questions. The prime functionalities of this proposed system is

- Identify Entities
- Identify Scope
- Identify relevant attributes
- Crawl what you need instead of what you get

The goal of this project is to index one million facts of different categories and test the system using different Natural Language queries. We choose the following most popular 20 different categories for this experiment.

- 1. Universities Data (1700 USA universities with statistics and reviews)
- 2. Countries (180 countries basic data)
- 3. Nutrition data (8560 food items)

- 4. Animals facts (800 different animals and species)
- 5. Movies (3500 English movies from the year 2000 - 2015)
- 6. Film personalities (13200 profiles)
- 7. Programming languages (PHP) (4000 different syntax and examples)
- 8. Gadgets (laptops/ mobiles and tablets) (1000 models)
- 9. Automobile (400 Cars with full specifications)
- 10. Famous people (4500 profiles)
- 11. Local restaurants (500 restaurants in Wichita with menu)
- 12. Companies (2000 USA companies basic information)
- 13. Websites (Alexa top 1 lakh domains data)
- 14. Basic Expression solver (supports 11 different operations)
- 15. Meaning, Synonyms and Antonyms for WordNet (117000 words)
- 16. AccuWeather Data (51318 locations in USA)
- 17. Mathematical conversions (400 conversions of size, area and volume)
- 18. Currency Exchange prices
- 19. Stock data (to be indexed)

- 20. Basic General Knowledge (to be indexed)

4.1 System Design

There are three important parts in designing the system

4.1.1 Extraction

Many pages in the World Wide Web follow their own patterns while representing the information. Our first discussion is about the Wikipedia website. Wikipedia is a semi structured knowledge repository where the data is organized in semi structured way. Except few most of the data is represented in form of hyperlinks, tables, free text and images. We choose 60 most reputed websites to crawl for information which includes Wikipedia, USDA.gov, US news, etc.

Initially every web page is requested using a "http" request. We used the "simple_html_dom.php" for parsing the web page in order to find the different elements in a web page like title, description, headings and other important features,. The crawler follows the links from the current page to next page automatically and will keep doing useful information into the dataset. Later the information which obtained on the each page is stored in the respective table and in the respective database.

4.1.2 Storing the crawled data

Before starting a crawler, we need to design the database with expected Entities and Attributes information and prepare the table ready to store the crawled information. We used MySQL to store the information. The crawled data is in MySQL tables in their respective database.

4.1.3 Parsing user query

Parsing User query is one of the challenging tasks for any search/answer engines. We are following our own way of extracting entities from the free text submitted by the user as a query. The first step is to remove stop words from the given query like for, the, in, at, etc. The next step is trying to match each keyword in the query with existing available entities, if matched the next keyword would be treated as feature/attribute the user is looking for.

Example: The following is a very simple example which will help in understanding how query is parsed.

Query: "**What is the** address **of** Wichita State University?"

In this query, first we need to eliminate the stop words, the stop words in this query are what, is, the, and, of.

Now, we the resultant query is "address Wichita State University." These resultant keywords are tried to match against the list of available entities. The word "Wichita State University" is found in the table "Universities" than the system understands that the query is related to an university. Now, it looks for another keyword in the resultant query which is "address". In our given table, it tries to pull up all the records that matched with Wichita State University and r gives the address column information as answer.

4.1.4 User-interface

The user interface is designed using PHP and HTML; the pages are created dynamically on fly based on the user questions. The images are loaded from the appropriate websites without maintaining a duplicate copy.

4.2 Experiment

The following is a brief discussion on some domains which crawled for this experiment.

a. **Universities:** The crawler which we built crawled and indexed 5000 web pages for different sets of information like university details, ranking, reviews, etc. Nearly 35 different basic attributes are identified for every university like name, address, established year, president, etc. To sum this up, we have 1700*35 nearly 60000 different facts were collected and these facts were preserved using MySQL.

b. **Nutrition:** For nutrition information 11000 different web pages were crawled and indexed for 8756 different food varieties. For each food 40 different attributes/features are identified like name, image source, vitamin a, magnesium, etc. which implies we have 3, 40,000 facts for nutrition information.

c. **Cars:** For cars, we crawled and indexed 260 different models each car have 70 different attributes thereby we collected 260*70 20,000 facts about cars

d. **Animals:** For this section, we crawled 4000 different types of animals with 11 attributes each implies 44000 different facts.

e. **Movies:** For movies facts, we indexed 20000 different wikipedia pages for information about movies, actors, directors, producers, singers. The movies were indexed between the years 2000 to 2015. Each movie contains 20 different features implies 3000*20 6000 different facts.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter discusses about the various experiments we conducted and detailed comparisons of proposed system performance with other search/answer engines like Google, Bing and Wolfram Alpha.

5.1 Query Working Mechanism

In this section we compared and discussed about the mechanism of parsing some basic queries in a Search Engine and proposed Semantic Engine.

5.1.1 Query 1: *"fruits with highest vitamin a"*

For this query, Google provided "1,980,000" search results. In order to provide these many results first Google needs to crawl, index and rank them. Let's say one page size is 10 KB implies the data that needs to store these many web pages equals 19.8 GB. For all the similar queries like nutrition let's make an estimated index size of 20000000 unique web pages of 10 KB each results 200 GB. In order to crawl, index and process these many web pages in a normal computer it takes a time of at least 60 hours by processing 100 pages per second. When comes to processing of this data, it is almost near impossible to process this much data in a normal computer with a normal processing power because of its volume. So that's why creating and managing a search engine is extremely difficult for normal startups as it requires huge number of servers with high-end processing powers.

On the user side, still the user needs to go through at least 10-20 results to draw a conclusion on "which fruits have highest vitamin A." Generally the user may frame hundreds of different unique queries like “vegetables with highest fiber”, “spices with highest calcium and magnesium”, etc. For every query, a regular Search Engine needs to process some millions and billions of web pages to provide the best possible results. In addition to that it needs to compute so many factors like the frequency of query, domain reputation, keyword(s) location in the page (url, title, heading, etc.), back links to the webpage, etc.

On the other hand, Semantic Search engine works quite differently. It systematically and semantically crawls one fact only one time and train the system to completely understand the data which it was indexing by keeping the possible user questions in mind so we index only what is important in a semantic way. We indexed the nutrition information from the trusted website "US Department of Agriculture." We collected the nutritional information for 8000 different foods and deducted nearly 320 thousand facts from the data. With this data, user may pose any kind of question like we discussed above and this semantic answer engine has capability to answer these questions. The time taken for crawling and indexing is about 8 minutes (1000 times less when compared with traditional search engine). The memory it takes to hold this data is just 3.3 MB (1000000 times less when compared with traditional search engine).

On the user side, user will experience direct, straight answer to the query instead of links which may contains answers. This is a simple example which helps in understanding how Semantic Search reduces the size of index to more than one million times. In addition to this we will save lot of power, infrastructure and complexity by enhancing the user experience to the best possible extent.

5.1.2 Query 2: *"Cars with highest Trunk volume"*

If a user wants to know the list of cars with highest trunk volume and if he poses this question on search engine, a normal search engine eliminates the stop words from the query and tries to find the web pages by matching the list of keywords with the web pages. The search engine initially tries to match with the exact keywords, if it doesn't find a match then it repeats the search with different synonyms like biggest, largest, huge in place of highest The assumption of Google search is, it expects there exists a page on World Wide Web discussing about the user question which means that there should be a web page for every single user question which is almost highly impossible.

Now let's discuss how the same question is processed in the proposed semantic engine. The given query is tokenized first and the stop words were eliminated from the query then it tries to find the entity in the question. Later it tries to find the feature the user is looking for in his/her query like "trunk volume" If we compare the lookup time, Google needs to go through some billions of web pages which contains the user entered keywords and further needs to filter them based on its rank and other factors. In our proposed system, the system just needs to compare the user submitted entity against less than 1000 entities help in saving enormous amount of lookup time.

5.2 Search Quality Analysis

In this section we are going to compare and discuss about the quality of search results among different popular search engines. This experiment is conducted with 20 queries for every category. The proposed system is capable of answering more queries in the given scope when compared with other search/answer engines. The following table shows the percentage of direct answers obtained in different search/answer engines.

	Universities	Nutrition	Cars	Movies	Film Personalities	Restaurants
Google	55%	65%	30%	55%	55%	85%
Bing	35%	50%	15%	50%	50%	85%
Wolfram Alpha	20%	70%	15%	20%	35%	0
Proposed System	80%	95%	70%	75%	90%	100%

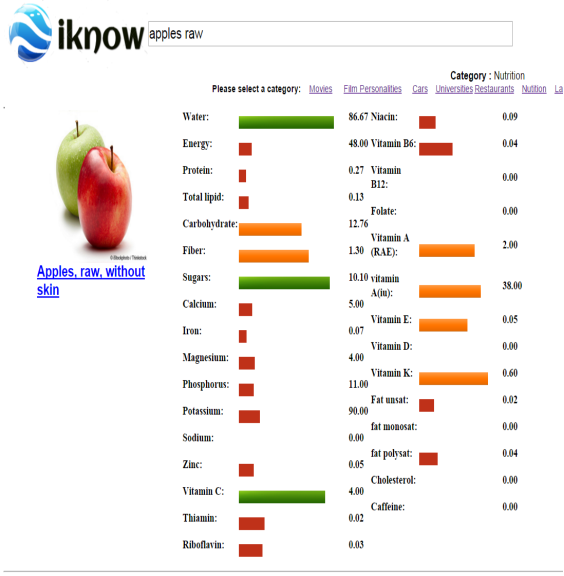
Table 5.1: Comparison of direct answers obtained in different search/answer engines

	Mobiles	Laptops	Animals	Countries	Instant answers	Web search
Google	20%	15%	70%	50%	100%	100%
Bing	15%	5%	55%	35%	100%	100%
Wolfram Alpha	10%	10%	10%	50%	95%	0
Proposed System	95%	80%	85%	95%	60%	30%

Table 5.2: Comparison of percentage accuracy in different search/answer engines

5.3. Comparison of Search Results:

The following are screenshots captured in 4 online search engines Google, bing, Wolfram Alpha and the proposed system (iknow.xyz). For the basic queries like nutrition information of a food item, all the four returned good informative results.



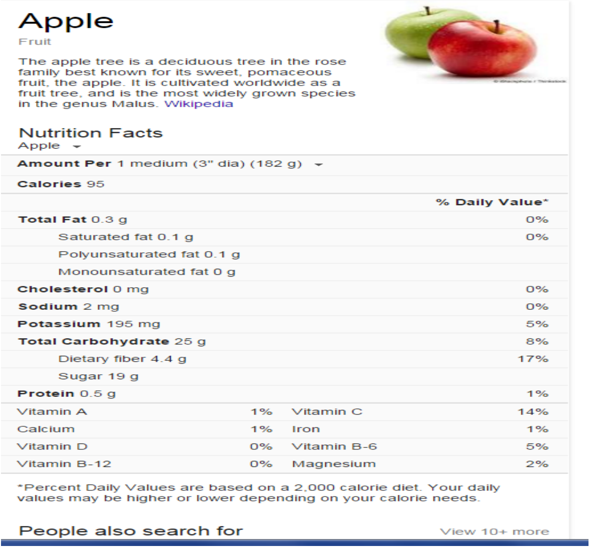
iknow apples raw

Category : Nutrition

Please select a category: Movies Film Personalities Cars Universities Restaurants Nutition La

Apples, raw, without skin

Water:	86.67	Niacin:	0.09
Energy:	48.00	Vitamin B6:	0.04
Protein:	0.27	Vitamin B12:	0.00
Total lipid:	0.13	Folate:	0.00
Carbohydrate:	12.76	Vitamin A (RAE):	2.00
Fiber:	1.30	vitamin A(iu):	38.00
Sugars:	10.10	Vitamin E:	0.05
Calcium:	5.00	Vitamin D:	0.00
Iron:	0.07	Vitamin K:	0.60
Magnesium:	4.00	Fat unsat:	0.02
Phosphorus:	11.00	fat monosat:	0.00
Potassium:	90.00	fat polysat:	0.04
Sodium:	0.00	Cholesterol:	0.00
Zinc:	0.05	Caffeine:	0.00
Vitamin C:	4.00		
Thiamin:	0.02		
Riboflavin:	0.03		



Apple

Fruit

The apple tree is a deciduous tree in the rose family best known for its sweet, pomaceous fruit, the apple. It is cultivated worldwide as a fruit tree, and is the most widely grown species in the genus Malus. Wikipedia

Nutrition Facts

Apple

Amount Per 1 medium (3" dia) (182 g)

Calories 95

	% Daily Value*
Total Fat 0.3 g	0%
Saturated fat 0.1 g	0%
Polyunsaturated fat 0.1 g	
Monounsaturated fat 0 g	
Cholesterol 0 mg	0%
Sodium 2 mg	0%
Potassium 195 mg	5%
Total Carbohydrate 25 g	8%
Dietary fiber 4.4 g	17%
Sugar 19 g	
Protein 0.5 g	1%

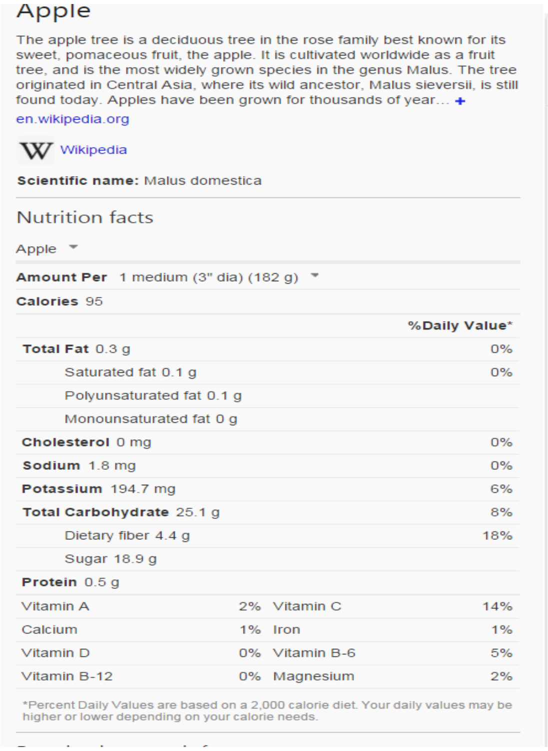
Vitamin A	1%	Vitamin C	14%
Calcium	1%	Iron	1%
Vitamin D	0%	Vitamin B-6	5%
Vitamin B-12	0%	Magnesium	2%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

People also search for View 10+ more

Fig. 5.1: screenshot of iknow.xyz(nutrition)

Fig. 5.2: screenshot of Google(nutrition)



Apple

The apple tree is a deciduous tree in the rose family best known for its sweet, pomaceous fruit, the apple. It is cultivated worldwide as a fruit tree, and is the most widely grown species in the genus Malus. The tree originated in Central Asia, where its wild ancestor, Malus sieversii, is still found today. Apples have been grown for thousands of year… +
en.wikipedia.org

W Wikipedia

Scientific name: Malus domestica

Nutrition facts

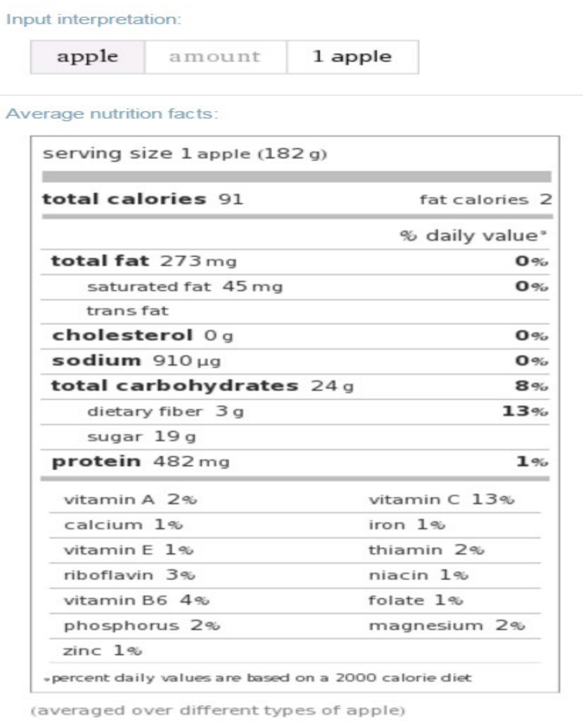
Apple ▾

Amount Per 1 medium (3" dia) (182 g) ▾

Calories 95

			%Daily Value*
Total Fat 0.3 g			0%
Saturated fat 0.1 g			0%
Polyunsaturated fat 0.1 g			
Monounsaturated fat 0 g			
Cholesterol 0 mg			0%
Sodium 1.8 mg			0%
Potassium 194.7 mg			6%
Total Carbohydrate 25.1 g			8%
Dietary fiber 4.4 g			18%
Sugar 18.9 g			
Protein 0.5 g			
Vitamin A	2%	Vitamin C	14%
Calcium	1%	Iron	1%
Vitamin D	0%	Vitamin B-6	5%
Vitamin B-12	0%	Magnesium	2%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.



Input interpretation:

apple	amount	1 apple

Average nutrition facts:

serving size 1 apple (182 g)

total calories 91 fat calories 2

			% daily value*
total fat 273 mg			**0%**
saturated fat 45 mg			**0%**
trans fat			
cholesterol 0 g			**0%**
sodium 910 µg			**0%**
total carbohydrates 24 g			**8%**
dietary fiber 3 g			**13%**
sugar 19 g			
protein 482 mg			**1%**
vitamin A 2%		vitamin C 13%	
calcium 1%		iron 1%	
vitamin E 1%		thiamin 2%	
riboflavin 3%		niacin 1%	
vitamin B6 4%		folate 1%	
phosphorus 2%		magnesium 2%	
zinc 1%			

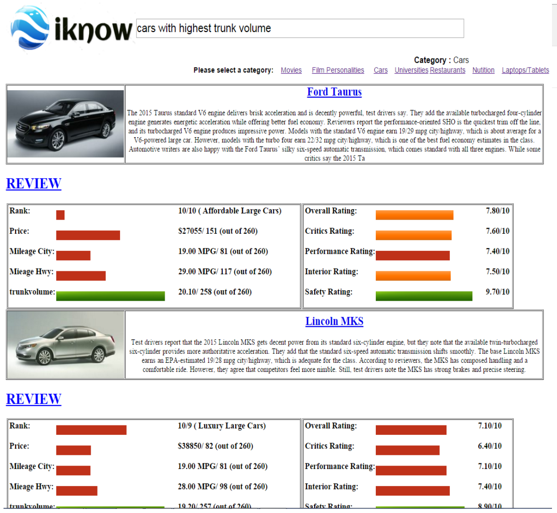
*percent daily values are based on a 2000 calorie diet

(averaged over different types of apple)

Fig. 5.3: screenshot of iknow.xyz(nutrition) Fig. 5.4: screenshot of Google(nutrition)

Next, we tried to find “cars with highest trunk volume” and for this search the proposed system (iknow.xyz) outperformed all the existing search engines. The figure 5.5 and figure 5.6 are the screenshots of results page in iknow.xyz and Google.

Next, we tried to find “celebs born los angeles” and for this search the proposed system (iknow.xyz) outperformed all the existing search engines. The figure 5.7 and figure 5.8 are the screenshots of results page in iknow.xyz and Google.



cars with highest trunk volume

Category : Cars

Please select a category: Movies Film Personalities Cars Universities Restaurants Nutition Laptops/Tablets

Ford Taurus

The 2015 Taurus standard V6 engine delivers brisk acceleration and is decently powerful, test drivers say. They add the available turbocharged four-cylinder engine generates energetic acceleration while offering better fuel economy. Reviewers report the performance-oriented SHO is the quickest trim off the line, and its turbocharged V6 engine produces impressive power. Models with the standard V6 engine earn 19/29 mpg city/highway, which is about average for a V6-powered large car. However, models with the turbo four earn 22/32 mpg city/highway, which is one of the best fuel economy estimates in the class. Automotive writers are also happy with the Ford Taurus' silky six-speed automatic transmission, which comes standard with all three engines. While some critics say the 2015 Ta

REVIEW

Rank:	10/10 (Affordable Large Cars)	Overall Rating:	7.80/10
Price:	$27055/ 151 (out of 260)	Critics Rating:	7.60/10
Mileage City:	19.00 MPG/ 81 (out of 260)	Performance Rating:	7.40/10
Mieage Hwy:	29.00 MPG/ 117 (out of 260)	Interior Rating:	7.50/10
trunkvolume:	20.10/ 258 (out of 260)	Safety Rating:	9.70/10

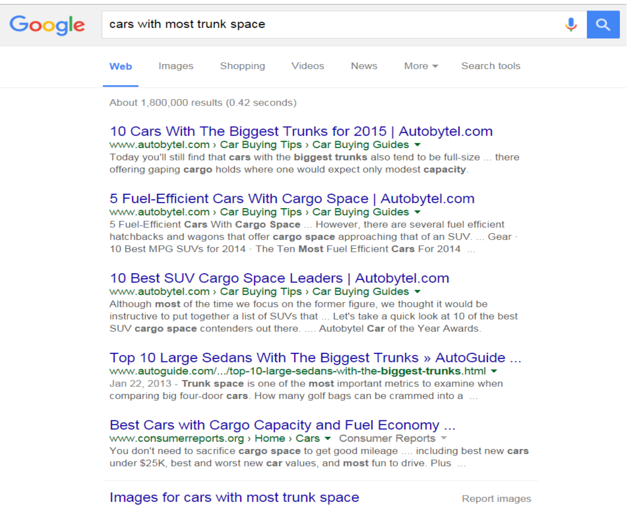
Lincoln MKS

Test drivers report that the 2015 Lincoln MKS gets decent power from its standard six-cylinder engine, but they note that the available twin-turbocharged six-cylinder provides more authoritative acceleration. They add that the standard six-speed automatic transmission shifts smoothly. The base Lincoln MKS earns an EPA-estimated 19/28 mpg city/highway, which is adequate for the class. According to reviewers, the MKS has composed handling and a comfortable ride. However, they agree that competitors feel more nimble. Still, test drivers note the MKS has strong brakes and precise steering.

REVIEW

Rank:	10/9 (Luxury Large Cars)	Overall Rating:	7.10/10
Price:	$38850/ 82 (out of 260)	Critics Rating:	6.40/10
Mileage City:	19.00 MPG/ 81 (out of 260)	Performance Rating:	7.10/10
Mieage Hwy:	28.00 MPG/ 98 (out of 260)	Interior Rating:	7.40/10
trunkvolume:	19.20/ 257 (out of 260)	Safety Rating:	8.90/10

Fig. 5.5: screenshot of iknow.xyz(cars)



cars with most trunk space

About 1,800,000 results (0.42 seconds)

10 Cars With The Biggest Trunks for 2015 | Autobytel.com
www.autobytel.com › Car Buying Tips › Car Buying Guides
Today you'll still find that **cars** with the **biggest trunks** also tend to be full-size ... there offering gaping **cargo** holds where one would expect only modest **capacity**.

5 Fuel-Efficient Cars With Cargo Space | Autobytel.com
www.autobytel.com › Car Buying Tips › Car Buying Guides
5 Fuel-Efficient **Cars** With **Cargo Space** ... However, there are several fuel efficient hatchbacks and wagons that offer **cargo space** approaching that of an SUV. ... Gear · 10 Best MPG SUVs for 2014 · The Ten **Most** Fuel Efficient **Cars** For 2014 ...

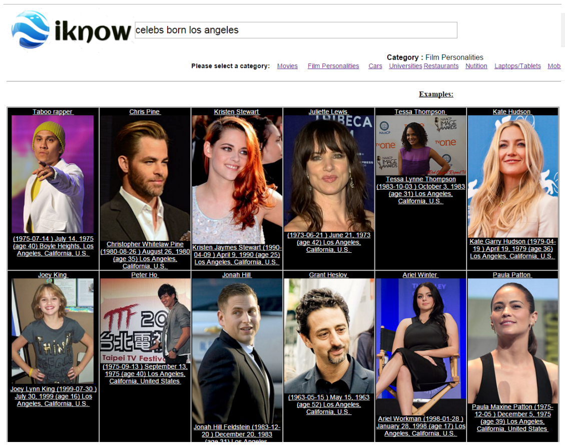
10 Best SUV Cargo Space Leaders | Autobytel.com
www.autobytel.com › Car Buying Tips › Car Buying Guides
Although **most** of the time we focus on the former figure, we thought it would be instructive to put together a list of SUVs that ... Let's take a quick look at 10 of the best SUV **cargo space** contenders out there. .... Autobytel **Car** of the Year Awards.

Top 10 Large Sedans With The Biggest Trunks » AutoGuide ...
www.autoguide.com/.../top-10-large-sedans-with-the-**biggest-trunks**.html
Jan 22, 2013 - **Trunk space** is one of the **most** important metrics to examine when comparing big four-door **cars**. How many golf bags can be crammed into a ...

Best Cars with Cargo Capacity and Fuel Economy ...
www.consumerreports.org › Home › Cars   Consumer Reports
You don't need to sacrifice **cargo space** to get good mileage .... including best new **cars** under $25K, best and worst new **car** values, and **most** fun to drive. Plus ...

Images for cars with most trunk space   Report images

Fig. 5.6: screenshot of Google(cars)



celebs born los angeles

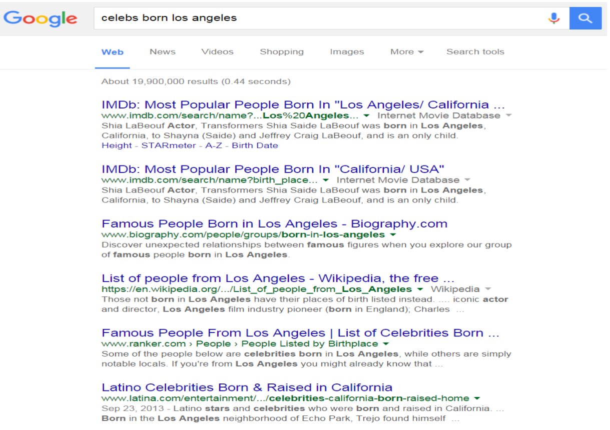
Category : Film Personalities

Please select a category: Movies Film Personalities Cars Universities Restaurants Nutition Laptops/Tablets Mob

Examples:

Taboo rapper	Chris Pine	Kristen Stewart	Juliette Lewis	Tessa Thompson	Kate Hudson
(1975-07-14) July 14, 1975 (age 40) Boyle Heights, Los Angeles, California, U.S.	Christopher Whitelaw Pine (1980-08-26) August 26, 1980 (age 35) Los Angeles, California, U.S.	Kristen Jaymes Stewart (1990-04-09) April 9, 1990 (age 25) Los Angeles, California, U.S.	(1973-06-21) June 21, 1973 (age 42) Los Angeles, California, U.S.	Tessa Lynne Thompson (1983-10-03) October 3, 1983 (age 31) Los Angeles, California, U.S.	Kate Garry Hudson (1979-04-19) April 19, 1979 (age 36) Los Angeles, California, U.S.
Joey King	Peter Ho	Jonah Hill	Grant Heslov	Ariel Winter	Paula Patton
Joey Lynn King (1999-07-30) July 30, 1999 (age 16) Los Angeles, California, U.S.	(1975-09-13) September 13, 1975 (age 40) Los Angeles, California, United States	Jonah Hill Feldstein (1983-12-20) December 20, 1983 (age 31) Los Angeles	(1963-05-15) May 15, 1963 (age 52) Los Angeles, California, U.S.	Ariel Workman (1998-01-28) January 28, 1998 (age 17) Los Angeles, California, U.S.	Paula Maxine Patton (1975-12-05) December 5, 1975 (age 39) Los Angeles, California, United States

Fig. 5.7: screenshot of iknow.xyz(celebrities)



celebs born los angeles

About 19,900,000 results (0.44 seconds)

IMDb: Most Popular People Born In "Los Angeles/ California ...
www.imdb.com/search/name?...**Los%20Angeles**...   Internet Movie Database
Shia LaBeouf **Actor**, Transformers Shia Saide LaBeouf was **born** in **Los Angeles**, California, to Shayna (Saide) and Jeffrey Craig LaBeouf, and is an only child.
Height - STARmeter - A-Z - Birth Date

IMDb: Most Popular People Born In "California/ USA"
www.imdb.com/search/name?birth_place...   Internet Movie Database
Shia LaBeouf **Actor**, Transformers Shia Saide LaBeouf was **born** in **Los Angeles**, California, to Shayna (Saide) and Jeffrey Craig LaBeouf, and is an only child.

Famous People Born in Los Angeles - Biography.com
www.biography.com/people/groups/**born-in-los-angeles**
Discover unexpected relationships between **famous** figures when you explore our group of **famous** people **born** in **Los Angeles**.

List of people from Los Angeles - Wikipedia, the free ...
https://en.wikipedia.org/.../List_of_people_from_**Los_Angeles**   Wikipedia
Those not **born** in **Los Angeles** have their places of birth listed instead. .... iconic **actor** and director, **Los Angeles** film industry pioneer (**born** in England); Charles ...

Famous People From Los Angeles | List of Celebrities Born ...
www.ranker.com › People › People Listed by Birthplace
Some of the people below are **celebrities born** in **Los Angeles**, while others are simply notable locals. If you're from **Los Angeles** you might already know that ...

Latino Celebrities Born & Raised in California
www.latina.com/entertainment/.../**celebrities-california-born-raised-home**
Sep 23, 2013 - Latino **stars** and **celebrities** who were **born** and raised in California. ... **Born** in the **Los Angeles** neighborhood of Echo Park, Trejo found himself ...

Fig. 5.8: screenshot of Google(celebrities)

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

The semantic engine we developed is only to demonstrate how to reduce the resources to the greatest extent by indexing just "what we need instead of what we get from the web." The current product is designed by keeping the most popular and most common user queries in mind and it may sometimes lacks an important information user is looking for. At this level, it will not replace the existing search engines. We hope that by increasing the number of categories and including even more number of features for each category results this answer engine capable of answering at least 70-80% of user queries. Some enthusiastic people who are part in a research or some related fields likes to dig deeper into the web results for finding all possible information on internet for their work, this solution may not fit for that kind of searches.

6.2 Future Work

In future, new categories need to be added into this Semantic Search. The prototype that built didn't have features like spell checking and suggestions, it needs to be included in the future. The proposed system is mostly dedicated to answer factual questions but there are categories like web search needs to be included in the future.

REFERENCES

LIST OF REFERENCES

[1] Web Search Engine, (https://en.wikipedia.org/wiki/Web_search_engine) [Cited 11/2015]

[2] Google web search engine, (https://www.google.com/) [Cited 11/2015]

[3] Computational Knowledge Engine - WolframAlpha (http://www.wolframalpha.com/) [Cited 11/2015]

[4] Answer Engine (knowledge base and semantic search engine software) True Knowledge (evi) (http://evi.com) [Cited 11/2015]

[5] Google semantic-search (Knowledge Graph) (https://en.wikipedia.org/wiki/Knowledge_Graph) [Cited 11/2015]

[6] Collaborative knowledge base - Freebase (http://freebase.com) [Cited 11/2015]

[7] Linked structured database - Wikidata (http://wikidata.org) [Cited 11/2015]

[8] Imielinski, T.; Signorini, A.,"If You Ask Nicely, I will Answer: Semantic Search and Today's Search Engines,"Semantic Computing, 2009. ICSC '09. Published in: Semantic Computing, 2009. ICSC '09. IEEE International Conference.

[9] Unaidah Mohamed Kassim and Mahathir Rahmany, "Introduction to Semantic Search Engine"

[10] Jagendra Singh, "A Comparative Study between Keyword and Semantic Based Search Engines”

[11] Mohamed Morsey, Jens Lehmann, Sören Auer, Claus Stadler and Sebastian Hellmann, "DBpedia and the Live Extraction of Structured Data from Wikipedia"

[12] José Paulo Leal, Vania Rodrigues, and Ricardo Queiros, " Computing Semantic Relatedness using DBPedia Schloss"

[13] Michael Schuhmacher Simone Paolo Ponzetto, " Exploiting DBPedia for Web search results clustering"

[14] Adel Tahri and Okba Tibermacine, "DBpedia Based Factoid Question Answering System"

[15] Inference William Tunstall-Pedoe, "True Knowledge: Open-Domain Question Answering Using Structured Knowledge and Inference"

[16] Alcides Calsavara and Glauco Schmidt, "Semantic Search Engines "

[17] Anusree. Ramachandran, R.Sujatha, "Semantic search engine: A survey"

[18] G.Madhu, Dr.A.Govardhan and Dr.T.V.Rajinikanth, "Intelligent Semantic Web Search Engines: A Brief Survey"

[19] Anne Aula, “Query Formulation in Web Information Search”