Graduation Project proposal
2011- 2012

Enhancement of entities comparison and search user experience using Linked data and sentiment analysis

1.MOTIVATION :

When Users are going to buy a product or form a decision about choosing something they tend to compare entities in with common features
Usually every kind of users care more for some distinctive features than others and usually they prefer entities with the best in these distinctive features (less price , portability , ..etc)

1.1 The old model :

The old model that websites always used to implement to preview products and preview reviews about products
As in (google products, Rottentomatoes.com, IMDB) is by crawling the web for the specific topic data (Movies for example)
And saving those data into database
Text processing is done on those data for extracting features , statistics or even sentiment
But it's still saved as text
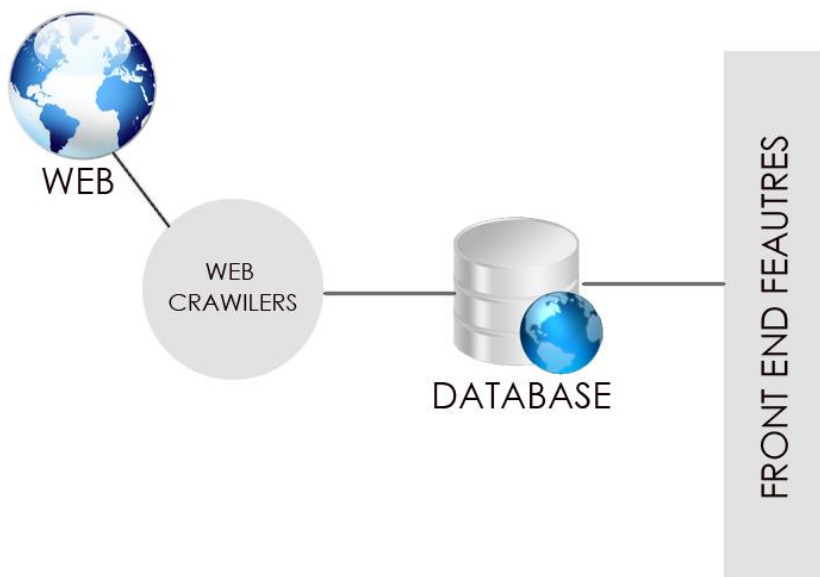
1.2 Old model problems:

- Collecting data is un-scalable and effortful
- Data is stored as Text and relations is only
- Entities common features are not linked together so u can't search for (hatch back red cars , restaurants in Egypt) unless a specific model is implemented to handle this Query which make kind of limitation

when searching for Hatch back cars it compares text similarities not features

1.3 project aim :

So our project mainly aims to enhance user experience when searching and comparison products and entities not depending on text matching but on features matching so Queries like HTC phones , Hatchback cars below $10000 can be done without a customized implementation in the System . As well as providing a detailed sentiment analysis on the product reviews to provide the community opinion in form of statistics to enhance comparison and ease decision taking

2.linked data :

2.1 - Linked data is a Good source for free available data:

As of March 2009
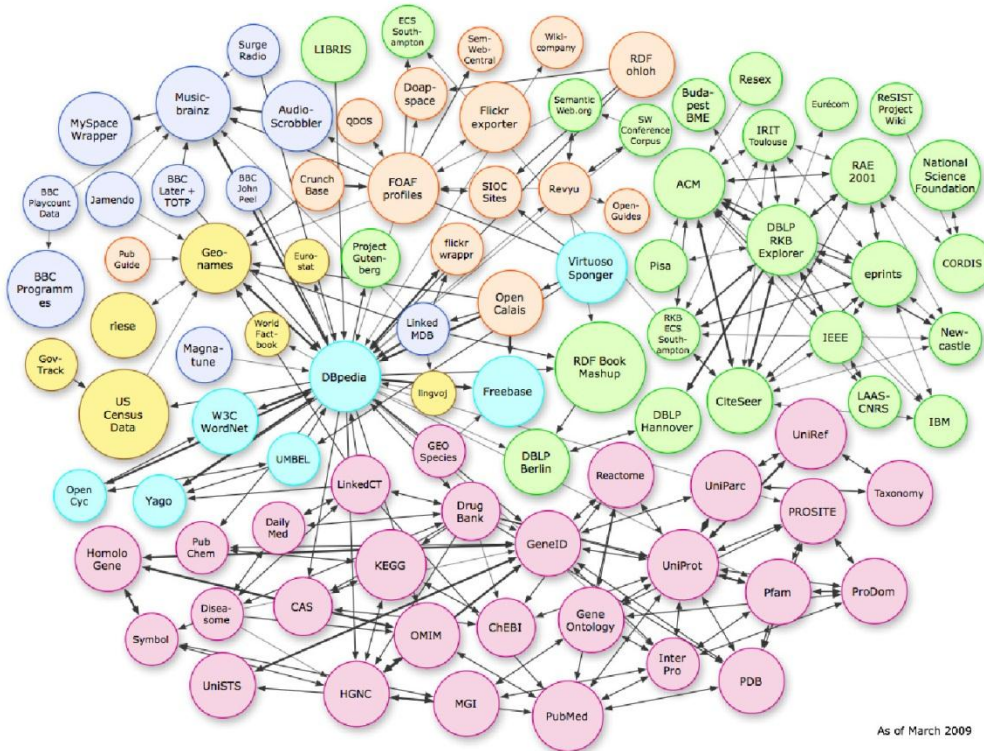
linked Data describes a method of publishing structured data so that it can be interlinked and become more useful. It builds upon standard Web technologies such as HTTP and URIs , but rather than using them to serve web pages for human readers

There's a lot of initiatives that offers a ready structured data that's available on the web , this data is offered as a linked data (Dbpedia is one of them) in which every entity , feature is an object and u can easily retrieve specific search Queries like (Ford red cars with type Hatch back , italian food restaurants in cairo) . Depending that type, location, color, vendor will be available as an objects in the linked data

Relation finder: http://www.visualdataweb.org/relfinder/relfinder.php   (a project based on linked data)

2.2 - How linked data will be useful in our case:

- ○ Effort saving & Scalability : linked data is already available and Growing unlike other websites that depends on scrapping more data from the web to add more products / entities , which need more infrastructure and more effort
- ○ Everything in linked and is an object : in linked data features are objects not text , for example (Barack Obama is a president of USA , and WDC is in united stated)

in both cases USA is the same entity so u don't have to match USA and united states through text processing

o Easily retrieving specific search result depending on comparing features objects not on text similarities (USA presidential candidates born outside USA , HTC Mobiles with Memory above 2 GB , Italian Food restaurants in Europe)

○ **The project architecture:**

SEARCH QUERY

FAST FOOD RESTAURANTS IN USA

WEB

SCRAPPERS

COLLECTING REVIEWS
AND STATISTICS OVER SPECIFIC ENTITIES

LINKED WITH LINKED DATA

RESULTS PAGE

DATABASE OF SPECIFIC STATISTICS
ON CERTIAN OBJECT
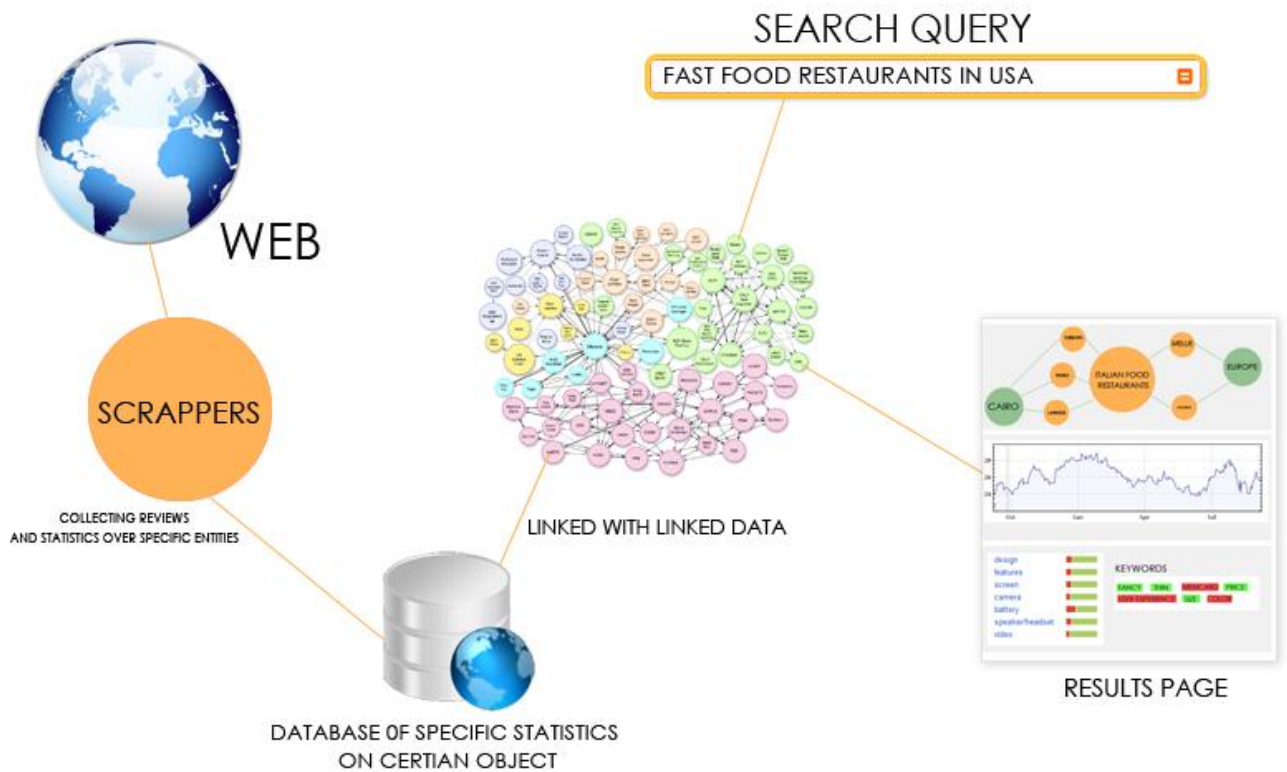
4. Sentiment analysis and feature extraction:

To provide new metrics to compare between entities public opinion is a good reliable feature to use to make your own opinion about a specific entity and rather than users have to read all reviews under each entity our project will provide 3 kind of statistics

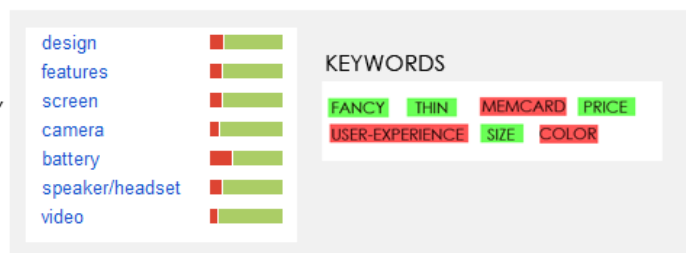LIST OF SEARCH RESULTS
VISUALIZED IN A GRAPH
WITH COMMON DISTICTIVE FEATURES

OVER ALL SENTIMENT ANALYSIS
OVER THE TIME LINE FOR ALL SEARCH RESULTS
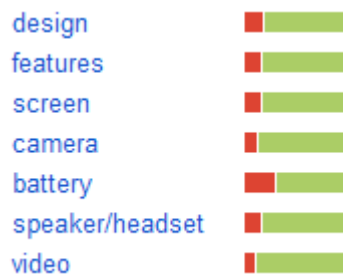OR THE CLICKED ENTITIY FROM THE GRAP

LIST OF SPECIFIC FEATURES FOR THE CLICKED ENTITY
SENTIMENT ANALYSISED ,
AND LIST OF EXTRACTED IMPOTANT KEYWORDS

- ○ Overall sentiment over the time line

○

- ○ Features extracted and sentiment of each Feature

- Extracted important Keywords that might contain High sentiment or have High frequency in the reviews

FANCY THIN MEMCARD PRICE
USER-EXPERIENCE SIZE COLOR

Different use cases :

- For collecting the Query it might take Different use cases :
  - Search for specific entity / product (Barack Obama)
  - Search for specific search Criteria (restaurants in Cairo , presidential candidates)
  - For a Given Examples get all similar cases (can be used for collaborative filtering , product recommendations , ads placement)

- **Some use-cases Examples :**
  - Calculating overall sentiment on Apple Products
  - Comparison between sentiment of presidential candidates in Egypt
  - Display sentiment for George Bosh over the time line
  - Comparison between Red Hatch back cars less than $10000
  - Deciding products that user might be interested in depending on user activity (recommendations , Ads placement)

Knowledge topics & Technology used :

- RDF / OWL to deal with linked data
- SparQl   to Query from linked data
- Sentiment analysis
- Natural language processing (NLTK is a good python tool in this case)
- Question answering
- Semantic search

○ Achievements so far :

the sentiment analysis part :

• we have explored the concept itself of sentiment analysis and different techniques used

• we have tried some techniques on a data set of bench mark annotated data of customer reviews which are :

  o Building a Dictionary and extracting TF-IDF vectors of the reviews and training a SVM Binary classifier to classify +ve and -ve reviews

  o second iteration was the same as above but using LSA on the extracted TF-IDF

  o we have used Python , and Gensim module for Text processing

• the above techniques was for classifying the whole sentence but for feature extraction and displaying sentiment over the each feature we have searched in some papers and sources , we found some techniques that could be doable like depending on part of speech taggers to extract the adjectives and relate them to the nearest noun . so far (so far stanford parser is a best example for a   natural language parser)

• we had Many chats and took opinions of people working in this domain and made a project based sentiment analysis and they were very useful to us and Guided us a lot

the semantic web and linked data part :

• well understanding of the Semantic web & linked data concepts through many articles / talks and videos / websites of W3S / papers

• we have spent some time in exploring different projects that is based on linked data and to how extend we can use it (some of the catchy ideas that took us it relation finder visualization project http://www.visualdataweb.org/relfinder/relfinder.php

• we understood the concepts of RDF / OWL through some tutorials and Documentations

• abstract understanding of how to Query the linked data using SparQL

• Exploration of LinkedData initiatives and what they offer like DBpedia

collecting data :

• we have explored the concepts of web Crawling and Scrapping Data from websites

• implemented a simple web scrapper using Python and Beautiful soap module

• having a previous experience in dealing with web APIs in Facebook and Twitter & those that could exist in reviews websites like Cnet & IMDB

Query interpretation :

• we did some exploration in this domain to ensure that this topic is applicable and we have reached that Query interpretation is Doable and have different techniques to implement it , but we didn't go in deep in any of those techniques