Project 2

Web Search Engine – Travelpedia

COEN 272 – Web Search and Information Retrieval
Department of Computer Science
Santa Clara University

Kavya Gautham W1166971 *kgautam@scu.edu*

Madhumitha Mani W1157297 *mmani@scu.edu*

Nisha Narayanaswamy W1153362 *nnarayanaswamy@scu.edu*

Priyanka Rao W1166161 *prao@scu.edu*

Abstract

The project aims at developing a robust web search engine for the travel domain. The targeted user base is anyone who is interested in finding vacation destinations within California. The search engine is built on ElasticSearch leveraging its various features. The goal of this project is to provide relevant search results to the users through various techniques like query boosting, query expansion, stemming, stopping, spell check, query parsing, handling geo-distance queries and automatic facet selection, which are in addition to the default features provided by Elasticsearch. The success of the web search engines: traditional and faceted, is determined by the effectiveness, efficiency and user satisfaction derived from the user evaluation studies.

1 Search Engine Overview

1.1 Introduction

The project is divided into five parts: (i) Crawl the travel related websites in California (ii) Build a traditional search engine using Elastic Search (iii) A non-traditional interaction feature using faceted search interface (iv) Logging component for user queries to log queries, clicks, session duration (v) Evaluation of the basic and advanced versions of the search engine through a user study. The architecture and the various features of the web search engine (Travelpedia) are detailed in the following sections.

1.2 High-level Architecture

The high-level architecture of the search is shown in the Figure 1.2. There are four stages in the development of Travelpedia – travel domain search engine (i) Crawling – getting the webpages related to the travel domain (ii) classifying/automatic labeling of the documents (iii) Indexing – involves creating mappers and analyzers for the documents (iv) Retrieval – process of retrieving documents from the search engine

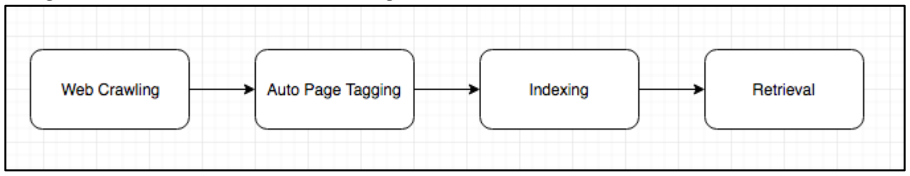


Figure 1.2 Search engine high-level architecture

2 Crawling and Document Indexing

2.1 Automatic document labeling

The automatic document labeling involves the following steps:

- Web pages from specific popular cities within California have been chosen and shortlisted for crawling. For example, the travel giants like travelpedia, visitcalifornia, discovercalifornia and many more specific sites for destinations throughout California were considered as the main seeds.

- A set of 400 specific sites covering all the facet categories were first chosen, crawled and manually labeled. Next, a set of 1300 more travel related pages were crawled and extracted for Auto Labeling process. The Auto Labeling process is described as follows:

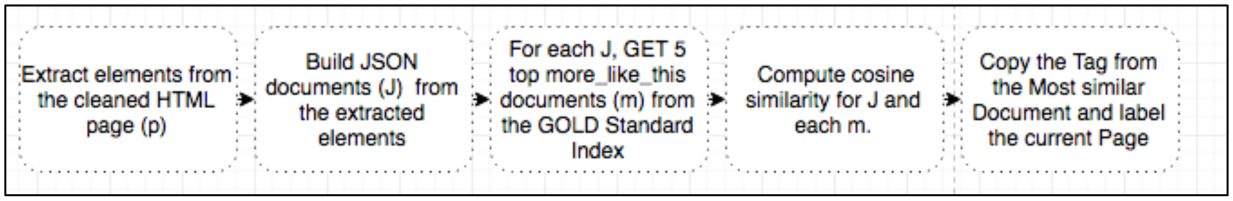

Extract elements from the cleaned HTML page (p)
Build JSON documents (J) from the extracted elements
For each J, GET 5 top more_like_this documents (m) from the GOLD Standard Index
Compute cosine similarity for J and each m.
Copy the Tag from the Most similar Document and label the current Page

Figure 2.1 Automatic document labeling

2.2 Document indexing

As part of pre-processing we do the following: when each HTML file is read, unwanted spaces and empty lines are removed. Then, the file is parsed through JSoup, to create the DOM structure. Once the DOM structure is created, script, comment, header, br, style, meta, CDATA tags are removed using JSoup, with the assumption that these tags will not provide any relevant information. Figure 2.2 shows the various steps involved before a document gets indexed into ElasticSearch.

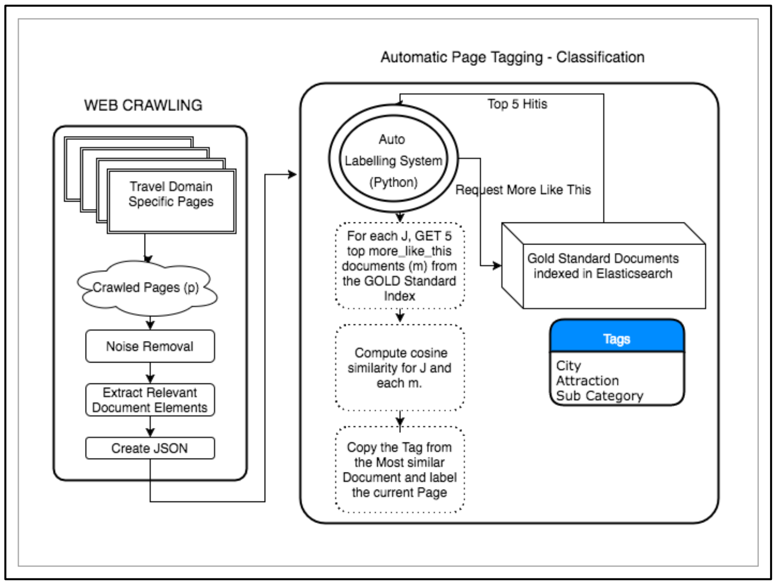

Automatic Page Tagging - Classification
WEB CRAWLING
Travel Domain Specific Pages
Crawled Pages (p)
Noise Removal
Extract Relevant Document Elements
Create JSON
Top 5 Hitis
Auto Labelling System (Python)
Request More Like This
For each J, GET 5 top more_like_this documents (m) from the GOLD Standard Index
Gold Standard Documents indexed in Elasticsearch
Tags
City
Attraction
Sub Category
Compute cosine similarity for J and each m.
Copy the Tag from the Most similar Document and label the current Page

Figure 2.2 Document indexing

The technique is based on the assumption all the viewable parts in a webpage are contained in the <body> tag. All the new metrics calculated were added as attributes to the nodes hence, keeping track of the metrics and making decisions per-node level was easy.

Before the documents are indexed, custom analyzers and mapping were created. Custom analyzers were created for stemming (Porter stemmer), stopping (English stopwords removed), lowercase, trimming (to remove extra spaces). The mapping uses the created custom analyzers for the various fields accordingly.

3 Document retrieval architecture

The search engine Travelpedia is built using Play framework. Play framework makes it easy to build web-applications with Java provides a lightweight, stateless, web-friendly architecture. We chose play framework as it provides a holistic platform for integrating the different components of the search engine like UI, logic processing component, interface to Elasticsearch. The following are the major components of the document retrieval model using Play framework:

3.1 View

The view is responsible for generating the HTML pages for user interaction with the search engine. There are two types of html pages generated: (i) Traditional Search Engine UI (ii) Faceted Search Engine UI. Both views provide user interface for login/logout component, executing search queries, viewing the SERP and pagination. In addition to these interfaces, faceted search engine UI has an additional component that enables users to apply filters for queries. Any action on the view elements triggers the execution of corresponding action in the controller.

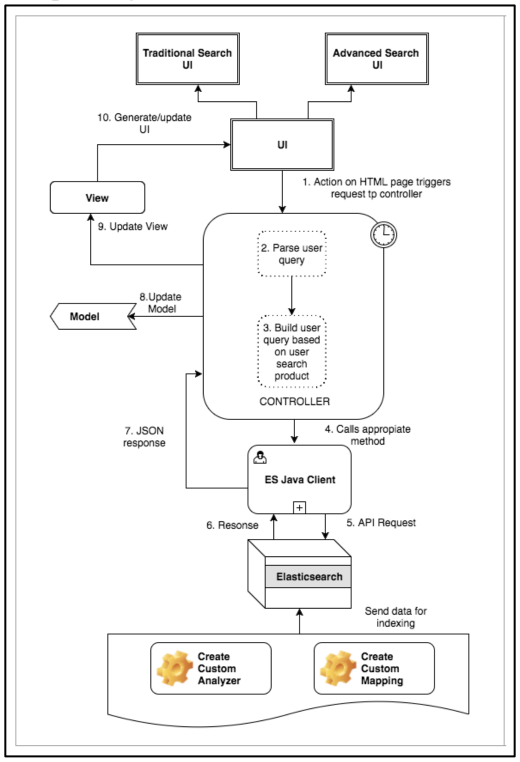


Figure 3.1 Document retrieval architecture

3.2 Controller

The controller is the crucial component of the document retrieval architecture as it wires together different components of the search engine. The controller handles all requests from the view page elements and populates the Model components necessary to perform search. Once all the relevant information is gathered, it invokes the corresponding service in the Java Client. Controller also waits to receive the response from the Java client which is in a JSON format, parse the JSON object to retrieve relevant results and populate the Model components that are necessary to render the SERP back to the user.

3.3 Java Client – Java interface to Elasticsearch

Elasticsearch is programming language independent and the APIs required for indexing, searching etc. could be accessed using HTTP and JSON in any language. Since the Travelpedia development team has strong Java expertise, it was a natural choice to use ElasticSearch Java client. The java client is created as a Transport client to perform API calls to ElasticSearch. It receives requests from the controller and converts it into JSON queries that are executed at the Elasticsearch. The response from the ElasticSearch is also in the JSON format, which is sent the controller that initiated this request.

3.4 Model

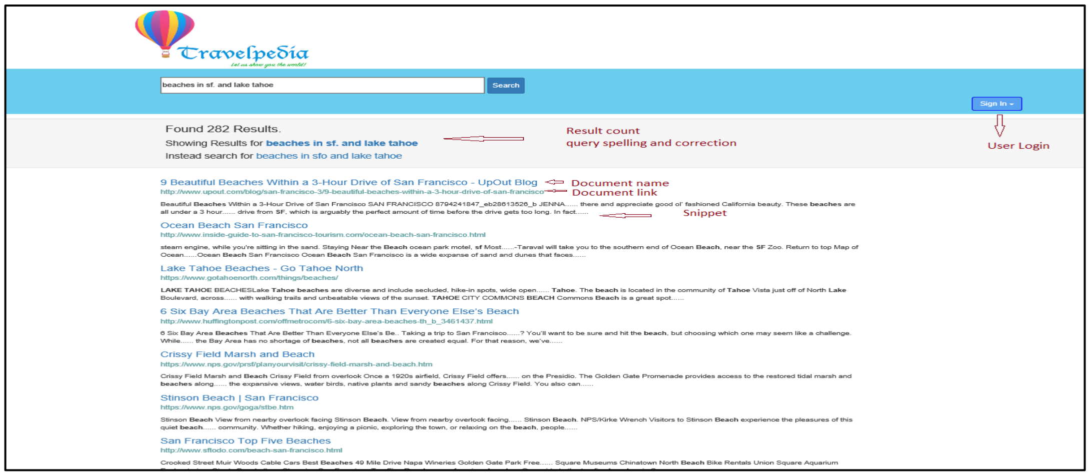
The model maintains the data needed to be passed across the requests and responses to and fro from UI and ElasticSearch. For example: the data retrieved from the search engine is stored as “SearchResult” objects, which contains a list of “Place” model objects. The Place model object is used to store search results returned from Elasticsearch like document title, liveURL, snippet and city information. Once the Controller populates the model objects, it also updates the View with the relevant information.

4 Versions of Search Engine

As mentioned earlier, the project demanded the development of two versions of search engines: the traditional and the advanced version. The underlying indexing and document retrieval architecture remains the same for both the search engines with just one striking difference in the advanced version, which is dealt in the following section.

4.1 Traditional Search Engine

The traditional search engine provides users with a simple interface to query documents related to travel destinations. The elements of the traditional SERP are **Input Box** - place to enter query, **Sign in button** - to login to the system, needed for user log, **Query correction** - replaces spelling error and provide suggestion, **Result count** - number of documents found for a particular query, **Title** - Document name, **Link** - url of the document, **Snippet** - small paragraph describing the document with query



beaches in sf. and lake tahoe

Found 282 Results.
Showing Results for **beaches in sf. and lake tahoe**
Instead search for beaches in sfo and lake tahoe

9 Beautiful Beaches Within a 3-Hour Drive of San Francisco - UpOut Blog
http://www.upout.com/blog/san-francisco-3/9-beautiful-beaches-within-a-3-hour-drive-of-san-francisco
Beautiful **Beaches** Within a 3-Hour Drive of San Francisco SAN FRANCISCO 8794241847_eb28613526_b JENNA...... there and appreciate good ol' fashioned California beauty. These **beaches** are all under a 3 hour...... drive from **SF**, which is arguably the perfect amount of time before the drive gets too long. In fact......

Ocean Beach San Francisco
http://www.inside-guide-to-san-francisco-tourism.com/ocean-beach-san-francisco.html
steam engine, while you're sitting in the sand. Staying Near the **Beach** ocean park motel, **sf** Most......-Taraval will take you to the southern end of Ocean **Beach**, near the **SF** Zoo. Return to top Map of Ocean......Ocean **Beach** San Francisco Ocean **Beach** San Francisco is a wide expanse of sand and dunes that faces......

Lake Tahoe Beaches - Go Tahoe North
https://www.gotahoenorth.com/things/beaches/
LAKE TAHOE BEACHESLake **Tahoe beaches** are diverse and include secluded, hike-in spots, wide open...... **Tahoe**. The **beach** is located in the community of **Tahoe** Vista just off of North **Lake** Boulevard, across...... with walking trails and unbeatable views of the sunset. **TAHOE** CITY COMMONS **BEACH** Commons **Beach** is a great spot......

6 Six Bay Area Beaches That Are Better Than Everyone Else's Beach
http://www.huffingtonpost.com/offmetrocom/6-six-bay-area-beaches-th_b_3461437.html
6 Six Bay Area **Beaches** That Are Better Than Everyone Else's Be.. Taking a trip to San Francisco......? You'll want to be sure and hit the **beach**, but choosing which one may seem like a challenge. While...... the Bay Area has no shortage of **beaches**, not all **beaches** are created equal. For that reason, we've......

Crissy Field Marsh and Beach
https://www.nps.gov/prsf/planyourvisit/crissy-field-marsh-and-beach.htm
Crissy Field Marsh and **Beach** Crissy Field from overlook Once a 1920s airfield, Crissy Field offers...... on the Presidio. The Golden Gate Promenade provides access to the restored tidal marsh and **beaches** along...... the expansive views, water birds, native plants and sandy **beaches** along Crissy Field. You also can......

Stinson Beach | San Francisco
https://www.nps.gov/goga/stbe.htm
Stinson **Beach** View from nearby overlook facing Stinson **Beach**. View from nearby overlook facing...... Stinson **Beach**. NPS/Kirke Wrench Visitors to Stinson **Beach** experience the pleasures of this quiet **beach**...... community. Whether hiking, enjoying a picnic, exploring the town, or relaxing on the **beach**, people......

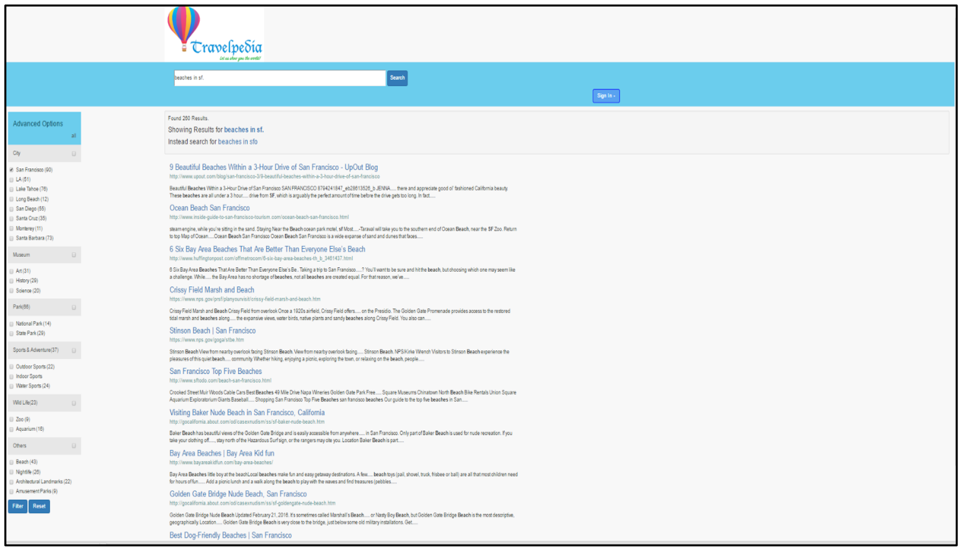
San Francisco Top Five Beaches
http://www.sftodo.com/beach-san-francisco.html
Crooked Street Muir Woods Cable Cars Best **Beaches** 49 Mile Drive Napa Wineries Golden Gate Park Free...... Square Museums Chinatown North **Beach** Bike Rentals Union Square Aquarium

Figure 4.1 Traditional SERP

word highlighted, **Pagination** - Result is divided into pages and each page displays 11 results. User can click next page number to view next page.

4.2 Faceted Search Engine

For the advanced version of the search engine, we decide to create facets for the following reasons: (i) there is an inherent hierarchy for the destinations like Parks which can be further classified as national/state park similarly Museums classified as Art, Science and History (ii) allows users to navigate easily according to several different topics than in a single, pre-determined order (iii) help users to learn new subcategories of the search. Figure 4.2 shows all the facets listed in our search engine.



beaches in sf.

Advanced Options
all

City
- [✓] San Francisco (90)
- LA (51)
- Lake Tahoe (76)
- Long Beach (12)
- San Diego (55)
- Santa Cruz (35)
- Monterey (11)
- Santa Barbara (73)

Museum
- Art (31)
- History (29)
- Science (20)

Park(66)
- National Park (14)
- State Park (29)

Sports & Adventure(37)
- Outdoor Sports (22)
- Indoor Sports
- Water Sports (24)

Wild Life(23)
- Zoo (9)
- Aquarium (16)

Others
- Beach (43)
- Nightlife (26)
- Architectural Landmarks (22)
- Amusement Parks (9)

Found 280 Results.
Showing Results for **beaches in sf.**
Instead search for beaches in sfo

9 Beautiful Beaches Within a 3-Hour Drive of San Francisco - UpOut Blog
http://www.upout.com/blog/san-francisco-3/9-beautiful-beaches-within-a-3-hour-drive-of-san-francisco
Beautiful **Beaches** Within a 3-Hour Drive of San Francisco SAN FRANCISCO 8794241847_eb28613526_b JENNA..... there and appreciate good ol' fashioned California beauty. These **beaches** are all under a 3 hour..... drive from **SF**, which is arguably the perfect amount of time before the drive gets too long. In fact.....

Ocean Beach San Francisco
http://www.inside-guide-to-san-francisco-tourism.com/ocean-beach-san-francisco.html
steam engine, while you're sitting in the sand. Staying Near the **Beach** ocean park motel, **sf** Most.....-Taraval will take you to the southern end of Ocean **Beach**, near the **SF** Zoo. Return to top Map of Ocean.....Ocean **Beach** San Francisco Ocean **Beach** San Francisco is a wide expanse of sand and dunes that faces.....

6 Six Bay Area Beaches That Are Better Than Everyone Else's Beach
http://www.huffingtonpost.com/offmetrocom/6-six-bay-area-beaches-th_b_3461437.html
6 Six Bay Area **Beaches** That Are Better Than Everyone Else's Be.. Taking a trip to San Francisco.....? You'll want to be sure and hit the **beach**, but choosing which one may seem like a challenge. While..... the Bay Area has no shortage of **beaches**, not all **beaches** are created equal. For that reason, we've.....

Crissy Field Marsh and Beach
https://www.nps.gov/prsf/planyourvisit/crissy-field-marsh-and-beach.htm
Crissy Field Marsh and **Beach** Crissy Field from overlook Once a 1920s airfield, Crissy Field offers..... on the Presidio. The Golden Gate Promenade provides access to the restored tidal marsh and **beaches** along..... the expansive views, water birds, native plants and sandy **beaches** along Crissy Field. You also can.....

Stinson Beach | San Francisco
https://www.nps.gov/goga/stbe.htm
Stinson **Beach** View from nearby overlook facing Stinson **Beach**. View from nearby overlook facing..... Stinson **Beach**. NPS/Kirke Wrench Visitors to Stinson **Beach** experience the pleasures of this quiet **beach**..... community. Whether hiking, enjoying a picnic, exploring the town, or relaxing on the **beach**, people.....

San Francisco Top Five Beaches
http://www.sftodo.com/beach-san-francisco.html
Crooked Street Muir Woods Cable Cars Best **Beaches** 49 Mile Drive Napa Wineries Golden Gate Park Free..... Square Museums Chinatown North **Beach** Bike Rentals Union Square Aquarium Exploratorium Giants Baseball..... Shopping San Francisco Top Five **Beaches** san francisco **beaches** Our guide to the top five **beaches** in San.....

Visiting Baker Nude Beach in San Francisco, California
http://gocalifornia.about.com/od/casexnudism/ss/sf-baker-nude-beach.htm
Baker **Beach** has beautiful views of the Golden Gate Bridge and is easily accessible from anywhere..... in San Francisco. Only part of Baker **Beach** is used for nude recreation. If you take your clothing off....., stay north of the Hazardous Surf sign, or the rangers may cite you. Location Baker **Beach** is part.....

Bay Area Beaches | Bay Area Kid fun
http://www.bayareakidfun.com/bay-area-beaches/
Bay Area **Beaches** little boy at the beachLocal **beaches** make fun and easy getaway destinations. A few..... **beach** toys (pail, shovel, truck, frisbee or ball) are all that most children need for hours of fun...... Add a picnic lunch and a walk along the **beach** to play with the waves and find treasures (pebbles.....

Golden Gate Bridge Nude Beach, San Francisco
http://gocalifornia.about.com/od/casexnudism/ss/sf-goldengate-nude-beach.htm
Golden Gate Bridge Nude **Beach** Updated February 21, 2016. It's sometimes called Marshall's **Beach**..... or Nasty Boy **Beach**, but Golden Gate Bridge **Beach** is the most descriptive, geographically Location..... Golden Gate Bridge **Beach** is very close to the bridge, just below some old military installations. Get.....

Best Dog-Friendly Beaches | San Francisco

Figure 4.2 Faceted SERP

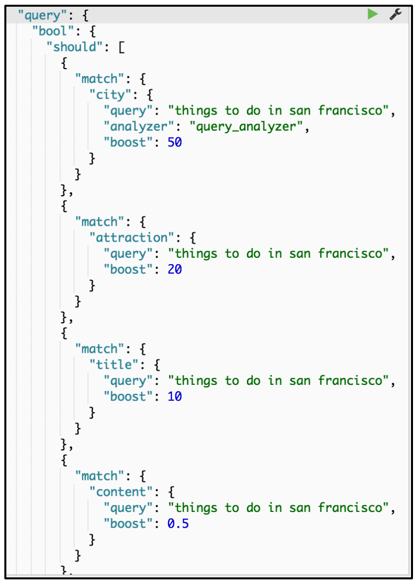
User can select one facet at a time or can select multiple facets using check-all checkbox, which selects group of facets. User can view the results of faceted query after clicking filter button. User can use reset button to unselect the facets selected in earlier query. There is no limit on facet selection. The results of the faceted query are displayed on right side of the same page. Each time a query is entered, the document counts of all facets are retrieved and displayed in facet drop down box.

5. General search engine features

There are a number of features that are common to both search engines as mentioned in the following section. The features were kept common as a design choice because we did not want to introduce bias towards any system.

5.1 Query-time Boosting

Boosting is an important tool used to tune relevance. Elastic search offers two types of boosting: (i) Index boosting (ii) Query-time boosting. In our case, index boosting was not necessary, as we had created only one index for all Travelpedia related documents. Figure 5.1 shows the snapshot of the query with boosting. Practically, there is no simple formula for specifying the "perfect" boost value for a query clause. It is a try-it-and-see task and finally depends on the design choice of the developers.



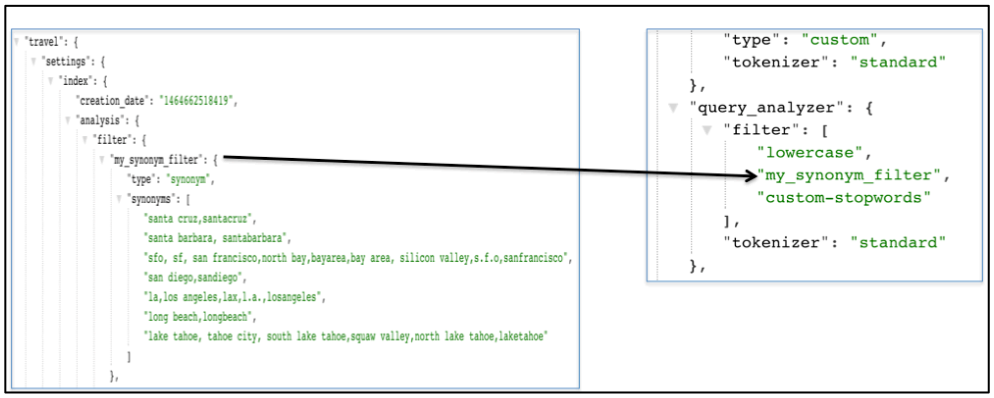
```
"query": {
  "bool": {
    "should": [
      {
        "match": {
          "city": {
            "query": "things to do in san francisco",
            "analyzer": "query_analyzer",
            "boost": 50
          }
        }
      },
      {
        "match": {
          "attraction": {
            "query": "things to do in san francisco",
            "boost": 20
          }
        }
      },
      {
        "match": {
          "title": {
            "query": "things to do in san francisco",
            "boost": 10
          }
        }
      },
      {
        "match": {
          "content": {
            "query": "things to do in san francisco",
            "boost": 0.5
          }
        }
      },
```

Figure 5.1 Query with boosting

ElasticSearch inherently supports "natural" boosting based on the "field-length norm". In Travelpedia, there was clearly a difference in the importance over various fields. For example: if the query terms matched against a city, it was boosted by a factor of 50, attraction and sub-category by a factor of 20 followed by title boosted by a factor of 10, keywords by 5 and finally the content by 0.5. This combination of boost values helped us achieve an acceptable level of document relevance across different query varieties.

5.2 Query Expansion

In most documents, the same concept may be referred using different words. In Travelpedia, for example: A user might search for "beaches in LA" and expect to find documents that contain "los angeles, la, losangeles, L.A." This issue is known as "synonymy" and has an impact on the overall recall of the information retrieval system. We overcome this problem only at the "city" level by introducing "synonyms" for the various cities we have at the time of indexing.



```
"travel": {
  "settings": {
    "index": {
      "creation_date": "1464662518419",
      "analysis": {
        "filter": {
          "my_synonym_filter": {
            "type": "synonym",
            "synonyms": [
              "santa cruz,santacruz",
              "santa barbara, santabarbara",
              "sfo, sf, san francisco,north bay,bayarea,bay area, silicon valley,s.f.o,sanfrancisco",
              "san diego,sandiego",
              "la,los angeles,lax,l.a.,losangeles",
              "long beach,longbeach",
              "lake tahoe, tahoe city, south lake tahoe,squaw valley,north lake tahoe,laketahoe"
            ]
          },
```

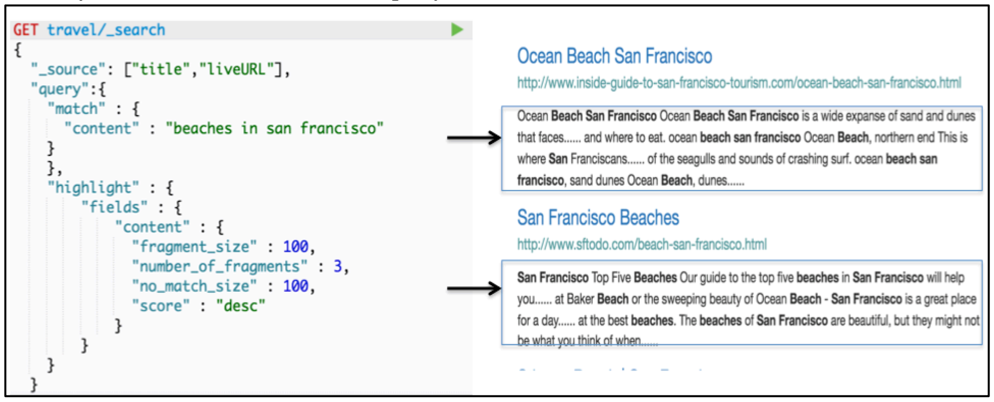
```
      "type": "custom",
      "tokenizer": "standard"
    },
    "query_analyzer": {
      "filter": [
        "lowercase",
        "my_synonym_filter",
        "custom-stopwords"
      ],
      "tokenizer": "standard"
    },
```

Figure 5.2 Query Analyzer with synonym for cities

As shown in Figure 4.2, query expansion is achieved using the concepts of "synonyms" in ElasticSearch. This helps expand the search query containing a city, into a number of queries with city variations based on the synonyms defined at the time of document indexing. This query expansion logic happens under the hood transparent to the user.

5.3 Snippet Generation and keyword highlight

Snippets are an important element of the SERP because it helps user get a sneak-peek of the document contents and help determine document relevance. Even more sophisticated snippets show "highlights" of text from the search result so a user can easily understand why the document matched the query.



```
GET travel/_search
{
  "_source": ["title","liveURL"],
  "query":{
    "match" : {
      "content" : "beaches in san francisco"
    }
    },
    "highlight" : {
      "fields" : {
        "content" : {
          "fragment_size" : 100,
          "number_of_fragments" : 3,
          "no_match_size" : 100,
          "score" : "desc"
        }
      }
    }
}
```

Ocean Beach San Francisco
http://www.inside-guide-to-san-francisco-tourism.com/ocean-beach-san-francisco.html

Ocean **Beach San Francisco** Ocean **Beach San Francisco** is a wide expanse of sand and dunes that faces...... and where to eat. ocean **beach san francisco** Ocean **Beach**, northern end This is where **San** Franciscans...... of the seagulls and sounds of crashing surf. ocean **beach san francisco**, sand dunes Ocean **Beach**, dunes......

San Francisco Beaches
http://www.sftodo.com/beach-san-francisco.html

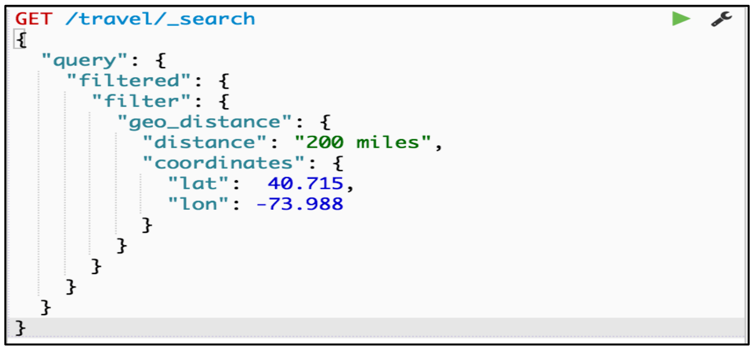
San Francisco Top Five **Beaches** Our guide to the top five **beaches** in **San Francisco** will help you...... at Baker **Beach** or the sweeping beauty of Ocean **Beach** - **San Francisco** is a great place for a day...... at the best **beaches**. The **beaches** of **San Francisco** are beautiful, but they might not be what you think of when......

Figure 5.3 Highlight query with generated snippets

Travelpedia achieves snippet generation with keyword highlighted using the "highlight" parameter in the search query. The highlight query code fragment and an example snippet are shown in Figure 5.3. The snippets are generated as fragments with a maximum of 3 fragments per snippet. Matching query keywords against the content field generates the snippet. Elasticsearch internally scores the fragments, and we use the top 3 fragments for display and to highlight the corresponding query keyword. If the keyword is not present in the content, first 100 words in the content will be printed.

5.4 Query Parsing to support geo-distance queries

When a user submits the search query, Travelpedia parses the query before any results are fetched from Elasticsearch. The aim of the query parser is to identity if the user has specified "distance-based" queries. For example: if the user is looking for state parks within 200 miles, the query parser identifies the distance as "200 miles" and narrows the search to places that are within the specified distance. The query and the corresponding results are shown in Figure 5.4



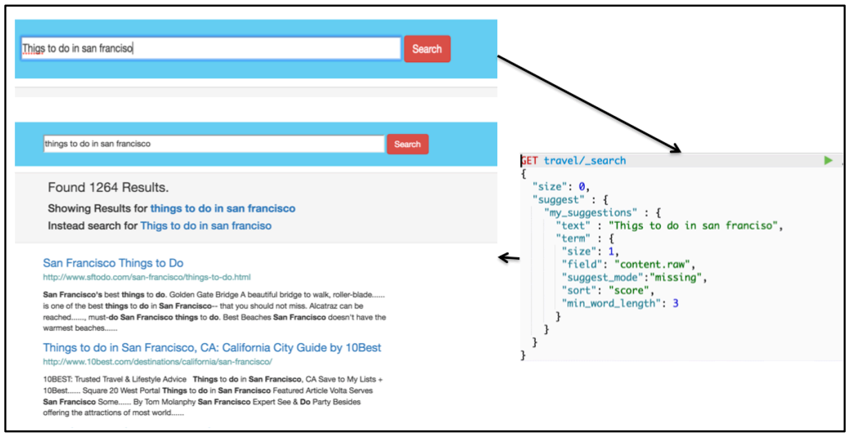
```
GET /travel/_search
{
  "query": {
    "filtered": {
      "filter": {
        "geo_distance": {
          "distance": "200 miles",
          "coordinates": {
            "lat":  40.715,
            "lon": -73.988
          }
        }
      }
    }
  }
}
```

Figure 5.4 Geo-distance query

If the query recognizes it as a distance-based query, the search is restricted to all the places that are within the specified distance. The user's IP address is captured and converted into geo-co-ordinates (latitude and longitude). The geo-coordinates are then added to the ElasticSearch query to get the appropriate results.

5.5 Spell check option

Similar to Google spell correction, Travelpedia also check for any spell-check errors in the user's search query. This feature leverages the "term" suggester API, which suggests terms based on edit distance.
The suggestions are given for those terms that are missing from the "content" field. For the purpose of spell check, we create another version of the content field where the data is not analyzed/stemmed/stopped. Travelpedia considers the top 1 correction suggested by Elasticsearch. One constraint for spell check is that the input data to term suggester should be a minimum of 3 characters in length. Figure 5.5 shows the example of the term suggester in action.



Found 1264 Results.

Showing Results for things to do in san francisco

Instead search for Thigs to do in san franciso

San Francisco Things to Do
http://www.sftodo.com/san-francisco/things-to-do.html
San Francisco's best things to do. Golden Gate Bridge A beautiful bridge to walk, roller-blade...... is one of the best things to do in San Francisco-- that you should not miss. Alcatraz can be reached......, must-do San Francisco things to do. Best Beaches San Francisco doesn't have the warmest beaches......

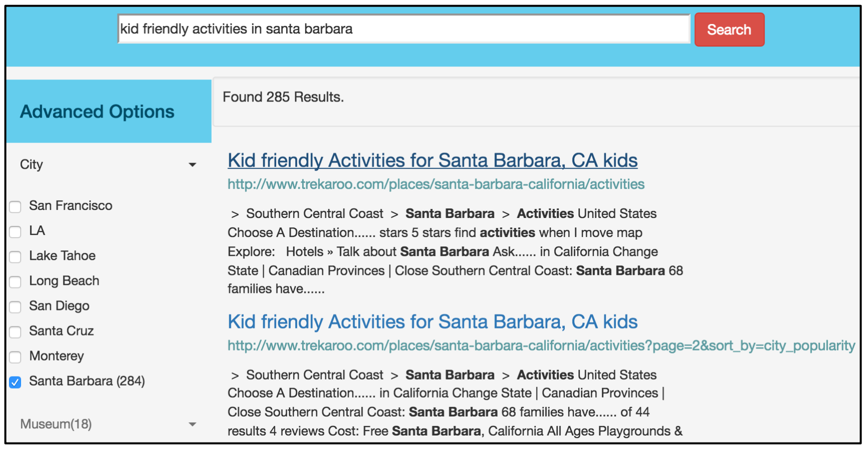
Things to do in San Francisco, CA: California City Guide by 10Best
http://www.10best.com/destinations/california/san-francisco/
10BEST: Trusted Travel & Lifestyle Advice Things to do in San Francisco, CA Save to My Lists + 10Best...... Square 20 West Portal Things to do in San Francisco Featured Article Volta Serves San Francisco Some...... By Tom Molanphy San Francisco Expert See & Do Party Besides offering the attractions of most world......

```
GET travel/_search
{
  "size": 0,
  "suggest" : {
    "my_suggestions" : {
      "text" : "Thigs to do in san franciso",
      "term" : {
       "size": 1,
       "field": "content.raw",
       "suggest_mode":"missing",
       "sort": "score",
       "min_word_length": 3
      }
    }
  }
}
```

Figure 5.5 Spell check example

5.6 Pre-selection of city facet

In the faceted search engine, if the user types cityname (which happens to one of the facets) in the search query, Travelpedia identifies it and pre-selects the city facet for the user. This is achieved by parsing the query, looking for possible city candidates that might be mentioned in the query. If the query parser returns any result, then we pre-select all the cities that the user has mentioned in the query.


kid friendly activities in santa barbara

Search

Advanced Options

Found 285 Results.

City

- [ ] San Francisco
- [ ] LA
- [ ] Lake Tahoe
- [ ] Long Beach
- [ ] San Diego
- [ ] Santa Cruz
- [ ] Monterey
- [✓] Santa Barbara (284)

Museum(18)

Kid friendly Activities for Santa Barbara, CA kids
http://www.trekaroo.com/places/santa-barbara-california/activities

> Southern Central Coast > **Santa Barbara** > **Activities** United States
Choose A Destination...... stars 5 stars find **activities** when I move map
Explore: Hotels » Talk about **Santa Barbara** Ask...... in California Change
State | Canadian Provinces | Close Southern Central Coast: **Santa Barbara** 68
families have......

Kid friendly Activities for Santa Barbara, CA kids
http://www.trekaroo.com/places/santa-barbara-california/activities?page=2&sort_by=city_popularity

> Southern Central Coast > **Santa Barbara** > **Activities** United States
Choose A Destination...... in California Change State | Canadian Provinces |
Close Southern Central Coast: **Santa Barbara** 68 families have...... of 44
results 4 reviews Cost: Free **Santa Barbara**, California All Ages Playgrounds &

Figure 5.6 Pre-selection of city facet

6 User Logging

For the purpose of search engine evaluation, we created logs of all the user activities across the search engines. We have created 5 user credentials for both traditional and faceted system. We are maintaining 5 user log files one for each user. Each user interaction is logged into his respective file.

6.1 Need for user logging

User logging helps us to understand the system from user perspective. For example: session time variable gives us an idea of time spent by each user on our system. The query list and pages viewed gives us an idea of which documents a user found more relevant for particular query and how many pages he had to click until he found his relevant document. These components can be used in future to improve the search engine. For example if many users viewed a subset of document for a particular query, we can boost such document up in the index so that next time these pages come as top results. User logging also helps in providing suggestion and recommendation to user.

6.2 Elements of the logs

Each time a user logs into the system we keep track of following components:

- Session variables (**start session time** - time at which user logged into the system; **end session time** - time at which user logged out)
- Query List - list of queries searched by a user in that session

```
Search Engine Type: Faceted SERP
Session start time: 2016/05/31 12:33:37
Queries: [boat lake tahoe, {cities: []  attraction: [Park]  subcategory: [National Park, State Park]},
                    {cities: []  attraction: [Park]  subcategory: [National Park, State Park]},
                    {cities: []  attraction: [Park]  subcategory: [National Park, State Park]},
                    {cities: [San Diego]  attraction: []  subcategory: []} ]
Mouse click count: 8
Pages_viewed : [1, 1, 3, 1, 2]
Session end time: 2016/05/31 12:37:32
```

Figure 6.1 User Logs

- Mouse clicks - number of links clicked for each user task
- Pages viewed - gives list of pages viewed for a particular query
- Engine Type - Traditional SERP or Faceted SERP
- For Faceted Search Engine: In addition to above variables we also keep track facet query search. Facet query search contains multiple facets selected by user to filter the result

6. User Evaluation

6.1 Evaluation goal

A small user study was conducted to compare the traditional and faceted version of the search engine. As part of usability research once both the search engine versions were developed, activities that a typical user would perform to achieve their goal on travel websites were listed. Five target participants between the ages 20-30 we identified. All were intermediate level users, including students and working professionals.

6.2 User study methodology

The procedures followed during the user study are as follows:

- Five target participants between the ages 20-30 we identified.
- Five tasks were identified to execute in both search engines.
- Task allocation matrix was created. Half the users were shown the traditional search engine to evaluate first followed by the faceted search engine. The other half was shown the faceted search first.
- A single system with Mac OSX was chosen for the users to evaluate on.
- Moderator provided initial introduction about the two search engine interfaces to users.
- Each user was given one task on each search engine interface. One user was evaluated at a time.
- Users were asked to login to the system. A new session was started before every task.
- Maximum time limit per task set as 10 minutes.
- Screen recording for all user evaluations, task duration was recorded using a digital stopwatch. The user logs were also recorded in each version of the search engine.
- Moderator did not provide any assistance to the users once the task evaluation started.
- Errors if any during evaluation were recorded.

A usability metric reveals some aspect of the user's interaction with the system. The metrics used for user evaluation were three-fold:

- **Efficiency** is the amount of effort required to complete a task. Measured the time taken, number of clicks, number of queries required completing each task. Since for each search engine, different users executed tasks 1 through 5, an average of each of the metrics across the two versions was calculated.
- **Effectiveness** is being able to complete a task. A task is successful if relevant documents were found accordingly to the user query. Three non-binary task success levels were defined and the user was asked to rate each task as "Success", "Partial Success" or "Failure". No assistance was provided to the user during task evaluation.
- **User satisfaction** determines how happy the user is with a given system. The USE questionnaire was provided to the user after a task was completed in each search interface. A five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" was used for rating.

6.3 User study results

The following five tasks were executed in both search engines in random order:

Task 1	You are visiting Monterey and would like to know the things you can do there. Try to narrow down to at least 3 activities of your choice.
Task 2	You are planning a bachelor/bachelorette party for your best friend in Los Angeles. You would like select a bar, restaurant for dinner.
Task 3	You want to plan a weekend trip to parks, museums and/or beaches within 200 miles.
Task 4	You would like to go whale watching. Pick a location of your choice, the best time and season to go.
Task 5	Plan your day in San Francisco

Figure 6.1: Tasks to be executed in traditional and faceted search engines

	TaskSet	TaskSet
User 1	SE1-Task1	SE2-Task2
User 2	SE2-Task3	SE1-Task2
User 3	SE1-Task3	SE2-Task4
User 4	SE2-Task5	SE1-Task4
User 5	SE1-Task5	SE2-Task1

Figure 6.2: Tasks allocation matrix

	Time per task *(in mins)*			Number of queries per task			Number of clicks per task	
User #	**SE1**	**SE2**		**SE1**	**SE2**		**SE1**	**SE2**
User 1	3.26	3.4		1	3		4	4
User 2	1.18	1.15		1	2		3	1
User 3	3.33	2.26		3	2		4	1
User 4	2.15	3.15		2	3		3	4
User 5	4.03	4.27		3	3		2	5
Average	**2.79**	**2.846**		**2**	**2.6**		**3.2**	**3**

Figure 6.2: Efficiency

	Task Success *(success/partial success/failure)*	
	SE1	**SE2**
User 1	Success	Success
User 2	Success	Success
User 3	Success	Partial Success
User 4	Success	Success
User 5	Partial Success	Success

Figure 6.3: Effectiveness

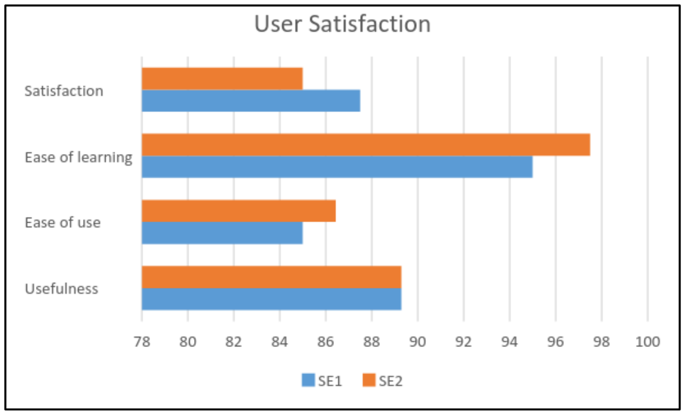


Figure 6.4: User Satisfaction Summary

7 Challenges faced

- Crawling only relevant documents and discarding the irrelevant
- Creating the gold standard manually as a baseline for auto-labeling of pages

- Ensuring the learning algorithm for automatic labeling improved with new incoming pages
- Coming up the robust architecture to handle end-to-end components of the search engine
- Understanding the Elasticsearch and leveraging the correct and relevant APIs
- Creating appropriate mapping and indexing for the documents

8 Conclusion and Future work

The user evaluation shows that both the search engines are on-par in terms of effectiveness, efficiency and user satisfaction. The future scope involves:

- Using query logs to boost search results
- Integrate with maps
- Query suggestions
- Expand city facets beyond California

References

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html

[2] https://www.playframework.com/documentation/2.5.x/JavaHome