Search Engine for Local and Global Businesses

Anonymous Author(s)
Affiliation
Address
email

Abstract

Many users now-a-days look for local businesses or places to dine at on the web. With access to a lot of data online, one can go online and search for businesses and read about not only the service provided by the business but also experiences of people who have interacted with them. We have built a search engine that makes it easier for one to find businesses locally and also read reviews and look at the ratings provided by their respective customers. We not only provide businesses that a user is looking for, but also recommend other such businesses that the user might like.

1 Motivation

Searching for businesses online has become easier now-a-days. Search engines such as Yelp's has provided users with information about businesses and also opinions of people who have used services provided by the business. Ratings and reviews go a long way in determining the success of a business and people invest considerable amount of time providing their experiences and suggestions as to how could one leverage the business' services to get a positive experience. Our main aim was to make use of this user-generated information to further enhance the information retrieved by our search engine.

We leveraged information about businesses and users to build a search engine that not only searches businesses queried by our users but also recommends similar businesses in and around the vicinity of their searches. This similarity is derived from the reviews provided for each business. We also provide recommendations based on user similarity. We used the Map-Reduce framework to get the user-based and the content based similarities. The subsequent sections provide detailed discussions on how we incorporated recommendations into our search engines. The search engine also performs location based personalization such that the retrieved businesses pertain to the region our search engine is being queried from.

We have built our search engine such that one could not only use it for searching for a specific business but also use free text to look for businesses that provide services which match the user's queries. For example, one could look for places that serve pizzas by querying *pizzas* on our search engine.

2 Design and Architecture

2.1 Document Processing

2.2 Indexing

Since, the users issue keyword based search queries, we build an index that acts a lookup for words that occur We read in 56,000 text files containing information about businesses along with

their reviews and other meta-data and stored it in the form of an inverted index using byte encoding to compress it. We built an inverted index of words occurring in the review text and score them based on the occurrences in the text.

2.3 Ranking

2.3.1 Ranking based on Title, Categories and Reviews

Ranking based on the terms contained in the review text is more likely to retrieve relevant businesses. The first thing we do while retrieving is to try and match the queries with the title of the businesses, since retrieval of a business matching the query issued should be fast and thus the title is weighed heavily. We also look for matches in the categories the business belongs to.

2.3.2 Ranking based on number of Reviews and Ratings

Including number of reviews of a business in our ranking definitely improves the retrieval for a search query. The idea behind including number of reviews is that if a business has more reviews, then certainly a lot of people have used it for its services and this would further enhance the business' credibility. Just because a business has more reviews does not make it a good business. Including a measure of how good the business is for it's services in terms of ratings would retrieve top businesses. Addition of ratings with the number of reviews adds to the improvement and now the retrieval is not just better in terms of whether the query is matched to the business but also the credibility of the business and how good the business is for it's services. Number of reviews can be equivalent to the number of views of a document on a web page.

3 Implementation

3.1 Data

We used Yelp's academic dataset to build our search engine. The dataset provides business and review information for nearly 56,000 businesses belonging to Canada UK and the United States. There were a total of 1.9 million reviews and around 100,000 tips. The information was provided in 3 different json files which were merged and written into one text file for each business. The following table shows .....:

1	Business Id
2	Name
3	Latitude and Longitude
4	Ratings
5	Address
6	Categories
7	Reviews
8	Tips

3.2 Recommendation using Map-Reduce framework

3.2.1 Algorithm

3.3 Ranking

In the previous section, we discussed the different features being used for ranking the businesses being retrieved by the engine. Intuitively one would think about using a linear ranking that generates scores based on different features, weighing them and add the scores linearly. But this doesn't improve the retrieval as will be discussed in the next section. Taking a product of the scores based on different features improves the retrieval to a certain extent.

The reason ranking is done multiplicatively is because if score from one of the features is high while the score from another is low, the overall score does not increase to a great extent. This way of

aggressively ranking the businesses results in retrieval of businesses that truly satisfy almost all the requirements.

$$\text{total score} = \frac{\text{(cosine score)x(score number of reviews)x(score ratings)}}{\text{distance}}$$

3.4 Querying

One can search businesses directly by simply putting in the name of the businesses. The engine returns the corresponding business the user was looking for along with similar businesses based on the similarity computed from term matches and similar businesses looked by other users.

The other way one could issue queries would be to find businesses that meet certain requirements like *Mexican Restaurants*. The engine retrieves businesses that match the term not only in the Categories and the titles, but also in the review text. The resulting retrieval is ranked based on Cosine similarity, location, number of reviews and ratings.

Geographic location based retrieval is the other part of the project, we focussed on. The engine not only retrieves businesses that match the query but also makes sure the businesses are situated in the region from where the query is issued.

4 Evaluation

5 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

5.1 Citations within the text

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard BIBTEX style `unsrt` produces this.) As to the format of the references themselves, any style is acceptable as long as it is used consistently.

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]", not "In our previous work [4]". If you cite your other papers that are not widely available (e.g. a journal paper under review), use anonymous author names in the citation, e.g. an author of the form "A. Anonymous".

5.2 Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[2]

5.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

[1] Sample of the first footnote

[2] Sample of the second footnote

Table 1: Data Description

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

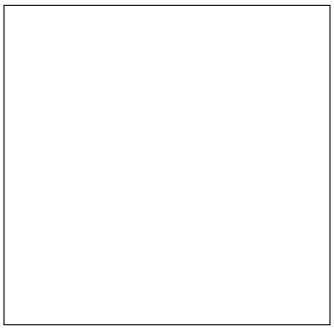

Figure 1: Sample figure caption.

5.4 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

6 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

7 Preparing PostScript or PDF files

Please prepare PostScript or PDF files with paper size "US Letter", and not, for example, "A4". The -t letter option on dvips will produce US Letter files.

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`

- LaTeX users:
  - Consider directly generating PDF files using `pdflatex` (especially if you are a MiKTeX user). PDF figures must be substituted for EPS figures, however.
  - Otherwise, please generate your PostScript and PDF files with the following commands:

    ```
    dvips mypaper.dvi -t letter -Ppdf -G0 -o mypaper.ps
    ps2pdf mypaper.ps mypaper.pdf
    ```

    Check that the PDF files only contains Type 1 fonts.
  - xfig "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
  - The `\bbold` package almost always uses bitmap fonts. You can try the equivalent AMS Fonts with command

    ```
    \usepackage[psamsfonts]{amssymb}
    ```

    or use the following workaround for reals, natural and complex:

    ```
    \newcommand{\RR}{I\!\!R} %real numbers
    \newcommand{\Nat}{I\!\!N} %natural numbers
    \newcommand{\CC}{I\!\!\!\!C} %complex numbers
    ```
  - Sometimes the problematic fonts are used in figures included in LaTeX files. The ghostscript program `eps2eps` is the simplest way to clean such figures. For black and white figures, slightly better results can be achieved with program `potrace`.
- MSWord and Windows users (via PDF file):
  - Select "Save or Publish to PDF" from the Office or File menu
- MSWord and Mac OS X users (via PDF file):
  - From the print menu, click the PDF drop-down box, and select "Save as PDF..."
- MSWord and Windows users (via PS file):
  - To create a new printer on your computer, install the AdobePS printer driver and the Adobe Distiller PPD file from `http://www.adobe.com/support/downloads/detail.jsp?ftpID=204` *Note:* You must reboot your PC after installing the AdobePS driver for it to take effect.
  - To produce the ps file, select "Print" from the MS app, choose the installed AdobePS printer, click on "Properties", click on "Advanced."
  - Set "TrueType Font" to be "Download as Softfont"
  - Open the "PostScript Options" folder
  - Select "PostScript Output Option" to be "Optimize for Portability"
  - Select "TrueType Font Download Option" to be "Outline"
  - Select "Send PostScript Error Handler" to be "No"
  - Click "OK" three times, print your file.
  - Now, use Adobe Acrobat Distiller or ps2pdf to create a PDF file from the PS file. In Acrobat, check the option "Embed all fonts" if applicable.

If your file contains Type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

7.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the graphicx package. Always specify the figure width as a multiple of the line width as in the example below using .eps graphics

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.eps}
```

or

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

for .pdf graphics. See section 4.4 in the graphics bundle documentation (`http://www.ctan.org/tex-archive/macros/latex/required/graphics/grfguide.ps`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to 'small' (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.