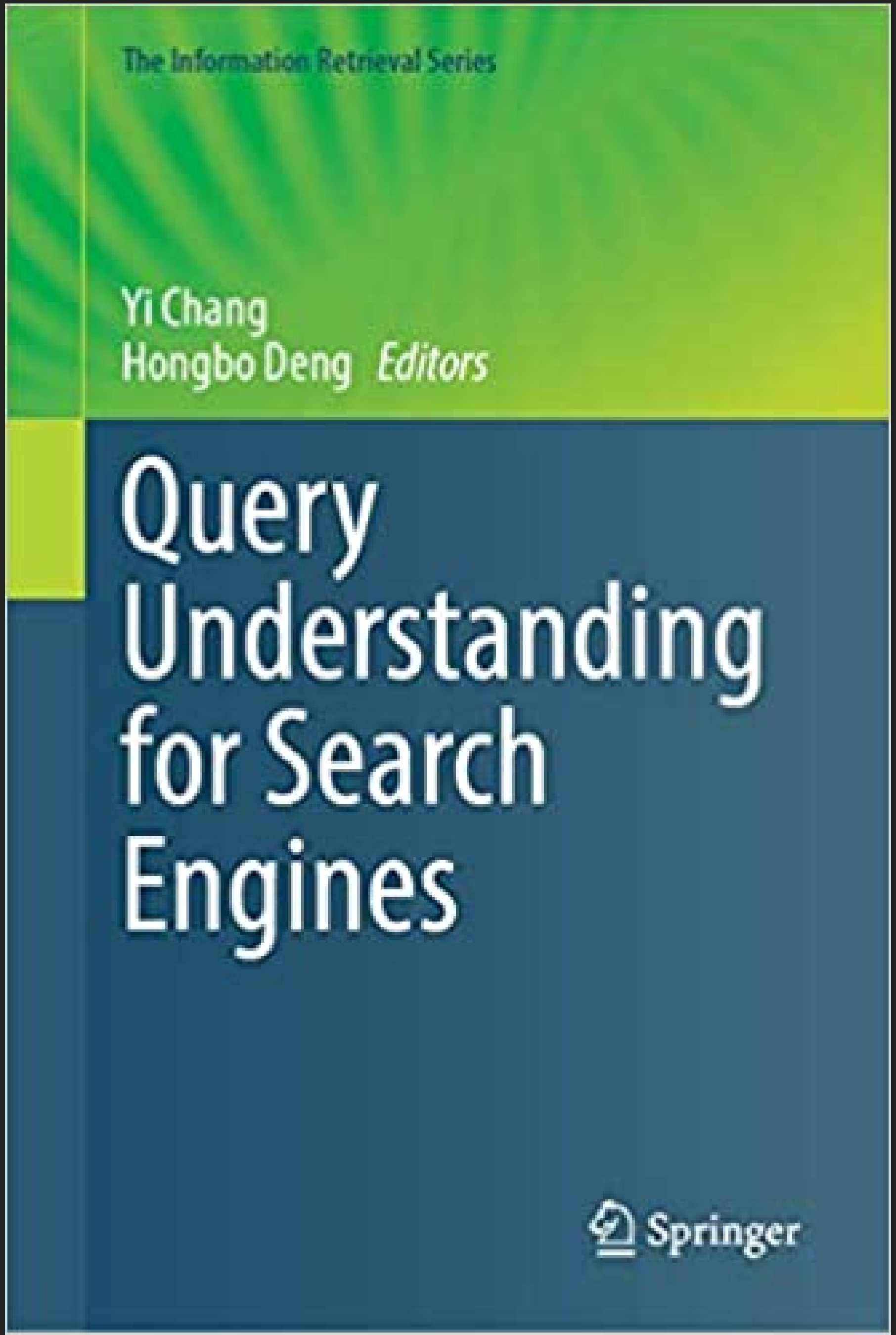
"Query Understanding for Search Engines"

chap.2 Query Classification

Speaker: @hurutoriya

Date: 2021-10-08

Book URL



The Information Retrieval Series

Yi Chang
Hongbo Deng Editors

Query Understanding for Search Engines

Springer

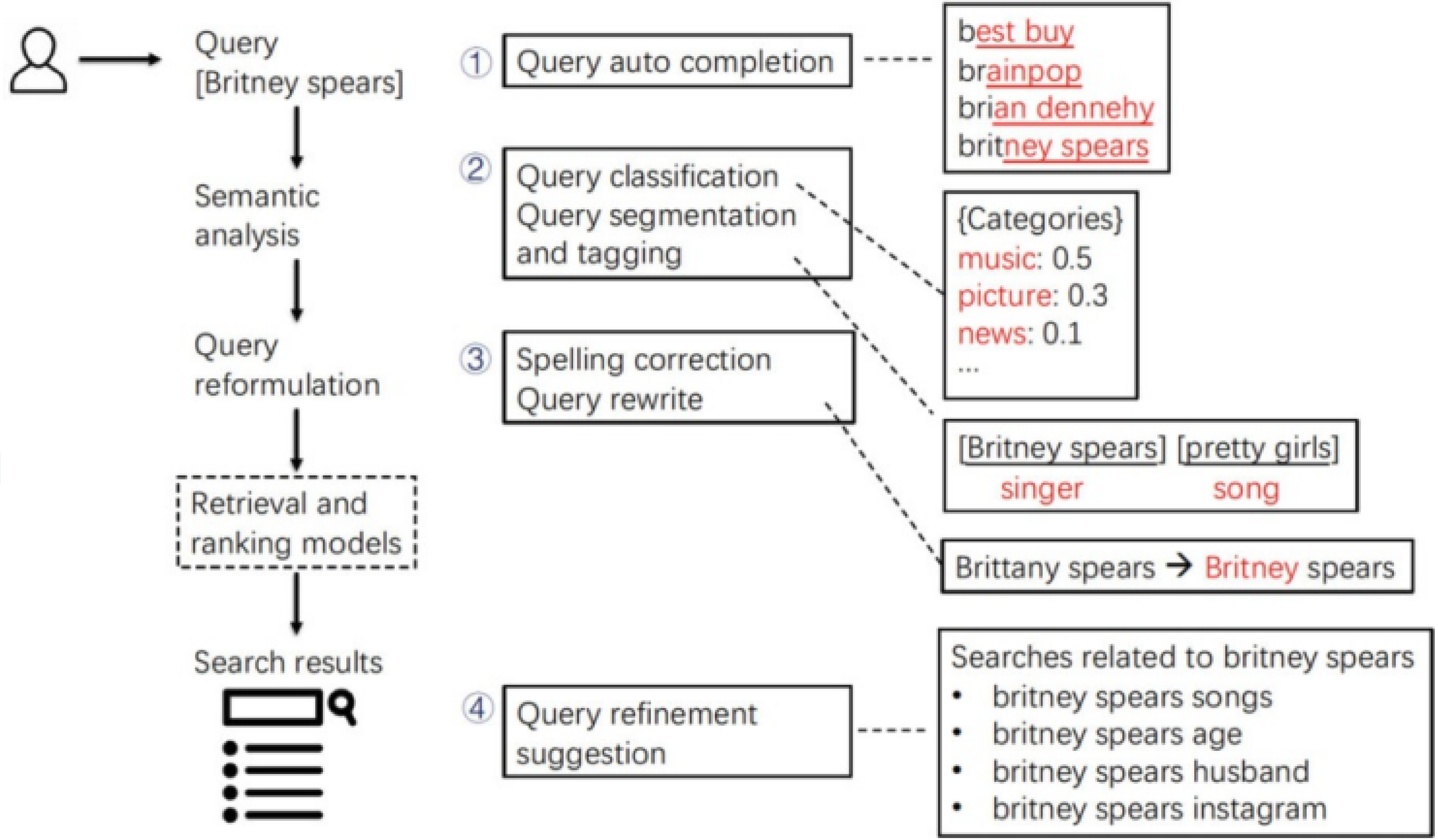


Fig. 1.2 An overview diagram of searching process

What is query Classification?

Definition

> Query classification, which is to assign a search query into a given target taxonomy.

e.g.

Query: "iPhone12"
↓
Query: "iPhone12" + Smartphone category (it is given category by your service)

Comparision to Unlike traditional document classification tasks

- It is much more difficult due to the
  - Short and ambiguous nature of queries
  - Demanding online computation requirement

Introduction

- Understanding what the user is searching for is at the heart of designing successful Web search applications

i.e., to assign a Web search query to one or more predefined categories.

Summarized 3 perspectives

1. Why: to understand customers’ search intent/goal—they might search to locate a particular site or to access some Web services
2. What(or Whern or Where): to understand search query’s topic, information type, geographic location, and time requirement
3. How: to understand how the search query performs—whether the results meet the curtomers’ expectations.

Group the existing works in query classification

- Intent Classification (Sec. 2.2)
- Topic Classification (Sec. 2.3)
- Performance Classification (Sec. 2.4)

Today we will talk until Topic Classification.

> Query classification, which is to assign a search query into a given target taxonomy.

- Original paper: A taxonomy of web search at SIGIR2002 by Broder.
- Refined Broder's paper: Understanding user goals in web search at WWW2004

classification methods

- manual classification:
  - Automatic identification of user goals in web search
  - Understanding user goals in web search at WWW2004
- automatic ones: using Decision tree and SVM.
  - The intention behind web queries
  - Determining the informational, navigational, and transactional intent of web queries
  - Query type classification for web document retrieval

Focus of proposing effective features for query intent identification.

Query Topic Classification

- It is critical to understand what the user is searching → It is usually very challenging.

why hard?: query is often highly vague, incomplete and subjective.

> If a search engine could successfully map search queries to some specific topics, the search results will be improved.

It could alleviate the ambiguity issues (e.g., jaguar the animal versus jaguar the car), by well capturing their topics.

Query topic classification is, therefore, defined to identify the underlying topics of queries according to some pre-defined topic taxonomy.

Query Topic Classification

- intermediate taxonomy for mapping (All papers at KDD2005)
  - The ferrety algorithm for the KDD cup 2005 problem
  - our winning solution to query classification in KDDCUP 2005
  - Classifying search engine queries using the web as background knowledge
- Robust classification of rare queries using web knowledge at SIGIR2007 focuse on Product saerch domain.

Topic Taxonomy

- KDD CUP-2005 Report: Facing a Great Challenge
- A formal two-level taxonomy, with 67 second level nodes and 800,000 internet user search queries

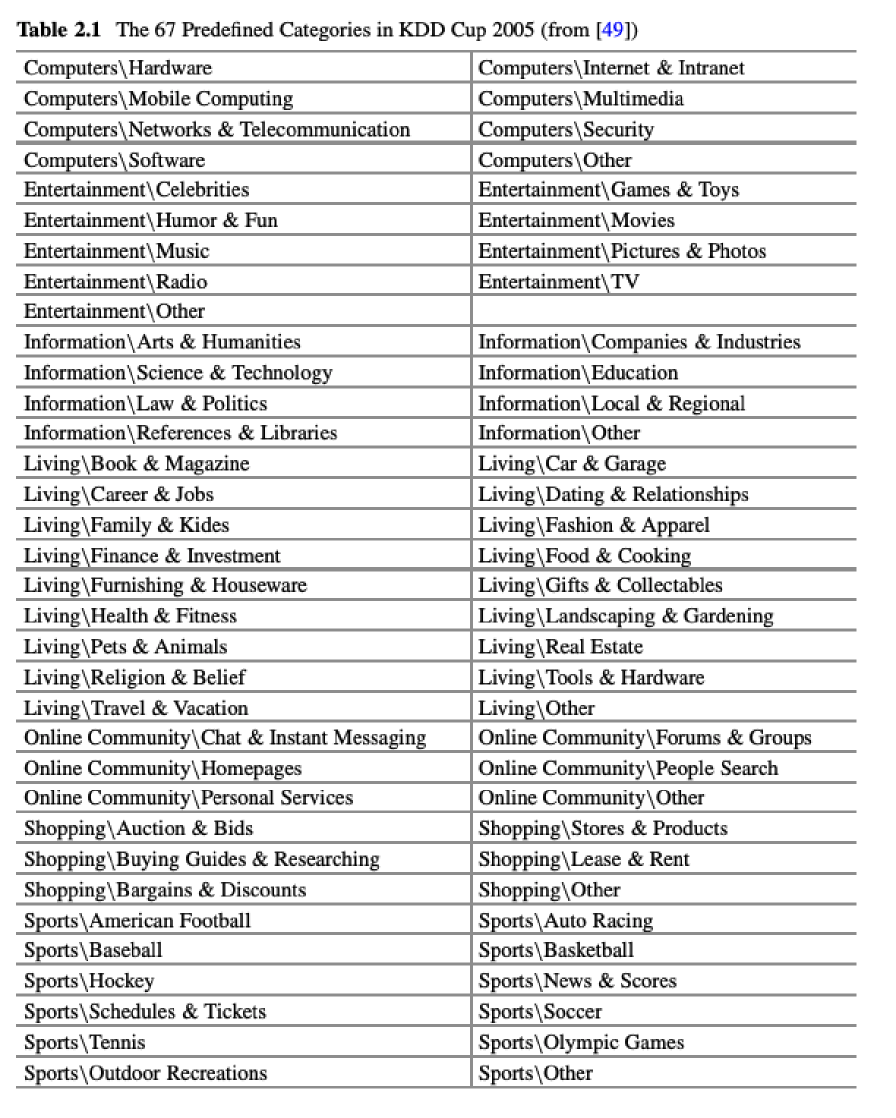


Table 2.1 The 67 Predefined Categories in KDD Cup 2005 (from [49])

Computers\Hardware	Computers\Internet & Intranet
Computers\Mobile Computing	Computers\Multimedia
Computers\Networks & Telecommunication	Computers\Security
Computers\Software	Computers\Other
Entertainment\Celebrities	Entertainment\Games & Toys
Entertainment\Humor & Fun	Entertainment\Movies
Entertainment\Music	Entertainment\Pictures & Photos
Entertainment\Radio	Entertainment\TV
Entertainment\Other	
Information\Arts & Humanities	Information\Companies & Industries
Information\Science & Technology	Information\Education
Information\Law & Politics	Information\Local & Regional
Information\References & Libraries	Information\Other
Living\Book & Magazine	Living\Car & Garage
Living\Career & Jobs	Living\Dating & Relationships
Living\Family & Kides	Living\Fashion & Apparel
Living\Finance & Investment	Living\Food & Cooking
Living\Furnishing & Houseware	Living\Gifts & Collectables
Living\Health & Fitness	Living\Landscaping & Gardening
Living\Pets & Animals	Living\Real Estate
Living\Religion & Belief	Living\Tools & Hardware
Living\Travel & Vacation	Living\Other
Online Community\Chat & Instant Messaging	Online Community\Forums & Groups
Online Community\Homepages	Online Community\People Search
Online Community\Personal Services	Online Community\Other
Shopping\Auction & Bids	Shopping\Stores & Products
Shopping\Buying Guides & Researching	Shopping\Lease & Rent
Shopping\Bargains & Discounts	Shopping\Other
Sports\American Football	Sports\Auto Racing
Sports\Baseball	Sports\Basketball
Sports\Hockey	Sports\News & Scores
Sports\Schedules & Tickets	Sports\Soccer
Sports\Tennis	Sports\Olympic Games
Sports\Outdoor Recreations	Sports\Other

Representative Work on KDD Cup Taxonomy

- Archived web site. you can download the dataset here
- there was no straight training data. KDD Cup 2005 only provided a small set of 111 queries with labeled categories→ not sufficient data size for supuervised learning...
    - participants can use other search engine, OSS to labeling the data. But...
    - Not explicit information about pre-defined topic-category.
    - Actuall dataset is noisy(miss spell)
    - Mannually categorize is impossible

Therefor need to design a scalable automatic classificaiton strategy.

KDD CUP-2005 report: facing a great challenge

- Preprocessing
    - Clean up noisy queries: stop words filltering, stemming and term frequency filtering
    - Advanced approach: spelling correction, compound word breaking, abbreviation expansion and named entity detection
- Gathering extra infomation
    - Motivation→ query is very short and hard to map the feature space or can not infer the meaning of query.
    - Another approach: augment queries. e.g. some participants used search result snippets, titles, and web pages to construct knowledge base, to expand query terms.

KDD CUP-2005 report: facing a great challenge

- Modeling: using SVM, KNN,Naive Baysian, LR and NN.
  1. directly mapped pre-defined directory structure to the target taxonomy, and produced required topics for each query.
  2. proposed to construct the mappings between the target topic categories and words or descriptions, so that some bag-of-words modeling strategies could be used to produce the categories of search queries.

Q2c@ust: our winning solution to query classification in KDDCUP 2005

- Phase I, they tackled the data sparsity problem by developing two kinds of base classifiers, a synonym-based classifier and a statistical classifier. Specifically, the synonym-based classifier was built by keyword matching between the enriched categories from search engine.
- tackle the feature sparsity problem, they used the search engine retrieved results to help represent a query, including the snippets, titles, URLs terms, and the category names in the directory.

Q2c@ust: our winning solution to query classification in KDDCUP 2005

- Phase II consisted of two stages. The first stage tackled the problem of lacking detailed query descriptions. Their strategy was to enrich queries by collecting their related web pages and category information through the use of multiple search engines, including Google and other search enginers.
- In the second stage, the enriched queries were then classified through the trained base classifiers trained