Visual interactive search and exploration of a digital library

Syed Sheece Raza Gardezi
Supervised by Hanna Suominan
u5752631
Contents

- Problem background
- Steps
- Related Work
- Road map
- Evaluation
- Conclusion
- Question Answer
Problem Background

- Data visualisation gives multiple dimensions to the human ability to perceive data.
- The number of published papers, in 23,750 journals, was an estimated 1.3 million in 2006. [1]
- Association for Computing Machinery (ACM), Elsevier, Institute of Electrical and Electronics Engineers (IEEE), and Springer, for example, are lacking in their capabilities to effectively visualize data.
- Ranked Lists no longer able to scale with the ever increasing amount of data.

Introduction to modern information retrieval
G Salton, MJ McGill - 1986 - citeulike.org
... acceptance of our use of cookies. You may hide this message. CiteULike is a free online bibliography manager. Register and you can start organising your references online. Tags. Introduction to Modern Information Retrieval ...
Cited by 13489 Related articles All 5 versions Import into BibTeX Save More

Information retrieval: data structures and algorithms
Abstract Information retrieval is a sub-field of computer science that deals with the automated storage and retrieval of documents. Providing the latest information retrieval techniques, this guide discusses Information Retrieval data structures and algorithms, ...
Cited by 2621 Related articles All 4 versions Import into BibTeX Save More

Term-weighting approaches in automatic text retrieval
G Salton, C Buckley - Information processing & management, 1988 - Elsevier
Abstract The experimental evidence accumulated over the past 20 years indicates that text indexing systems based on the assignment of appropriately weighted single terms produce retrieval results that are superior to those obtainable with other more elaborate text ...
Cited by 8319 Related articles All 18 versions Web of Science: 2272 Import into BibTeX Save More

Modern information retrieval
R Baeza-Yates, B Ribeiro-Neto - 1999 - mail.im.tku.edu.tw
Information retrieval (IR) has changed considerably in recent years with the expansion of the World Wide Web and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR textbooks have become a little out of date.

[PDF] cornell.edu
Find It at ANU

[PDF] tku.edu.tw
Homomorphic encryption for secure information retrieval from the cloud
V. Anand, Suresh Chandra Satapathy
2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)
Year: 2016
Pages: 1 - 5, DOI: 10.1109/ICETETS.2016.7602988
IEEE Conference Publications
Abstract (20371 Kb)

Agricultural information retrieval in geographically distributed networks
Chen Wei; Li Zhemin; Zhang Gao; Wang Dongjie; Wu Chen
2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)
Year: 2016
Pages: 1 - 4, DOI: 10.1109/Agro-Geoinformatics.2016.7577614
IEEE Conference Publications
Abstract (126 Kb)

Design of focused crawler for information retrieval of Indian origin Academicians
Manish Kumar; Rajesh Bhatia; Apoorva Ohri; Aditya Kohli
2016 International Conference on Advances in Computing, Communication, & Automation (iCACCA) (Spring)
Year: 2016
Pages: 1 - 6, DOI: 10.1109/iCACCA.2016.7578895
IEEE Conference Publications
Abstract (269 Kb)

Matrix Factorization-based clustering of image features for bandwidth-constrained information retrieval
Jacob Chakareski; Immanuel Manohar; Shantanu Rane
2016 IEEE International Conference on Multimedia & Expo Workshops
Year: 2016
Pages: 1 - 6, DOI: 10.1109/MMX.2016.7734438
IEEE Conference Publications
Abstract (285 Kb)
Towards a framework for attribute retrieval
Arifin Kopiliku, Mohand Beoughanem, Karen Pineau-Sauvagnat
October 2011 CIKM ’11: Proceedings of the 20th ACM international conference on Information and knowledge management
Publisher: ACM
Bibliometrics: Citation Count: 3
Downloads (6 Weeks): 9, Downloads (12 Months): 83, Downloads (Overall): 311

Full text available: PDF
In this paper, we propose an attribute retrieval approach which extracts and ranks attributes from HTML tables. We distinguish between class attribute retrieval and instance attribute retrieval. On one hand, given an instance (e.g., University of Strathclyde) we retrieve from the Web its attributes (e.g., principal, location, number of students)....
Keywords: attribute retrieval, information retrieval

Prototyping a personalized contextual retrieval framework
Damian Palacio, Guillaume Cabanac, Gilles Hubert, Karen Pineau-Sauvagnat, Christian Sallaberry
October 2013 GIR ’13: Proceedings of the 7th Workshop on Geographic Information Retrieval
Publisher: ACM
Bibliometrics: Citation Count: 0
Downloads (6 Weeks): 7, Downloads (12 Months): 40, Downloads (Overall): 69

Full text available: PDF
We introduce a framework for searching places according to user interests and spatial context. Our framework combines existing geo-tools or services (e.g., Google Places, Yahoo! BOSS Geo Services, PostGIS, Geophy, Geonames) and ranks results according to features such as distance, popularity, and user preferences. We used this framework to participate...
Keywords: contextual information retrieval, geographic information retrieval

Information filtering and information retrieval: two sides of the same coin?
Nicholas J. Belkin, W. Bruce Croft
Publisher: ACM
Bibliometrics: Citation Count: 266
Downloads (6 Weeks): 53, Downloads (12 Months): 360, Downloads (Overall): 7,284

Full text available: PDF
Information filtering and information retrieval are two sides of the same coin. Both strive to find relevant information from a large and dynamic collection of documents. Information filtering seeks to find information that is relevant to a user's needs, while information retrieval seeks to find information that is relevant to a query. However, the two fields have different goals and methods. Information filtering is concerned with finding the right document, while information retrieval is concerned with finding the right information within a document. In this paper, we will explore the relationship between the two fields and discuss how they can be used together to improve the quality of information retrieval.
Work Done

- Created a clean data set of
- Created a custom scoring function for solr
- Created a php-solr api for Solr 6.2 and php7
- Integrated Mallet, JATE, Weka
- Created a novel approach to label clusters
- Create a production ready example
- Created Visualisations to explore data
### Data

- **NICTA - 6000 papers**
- **Microsoft - 5000**
- **JLMR - 2500**
- **Google - 1300**

- **Clean data set**

<table>
<thead>
<tr>
<th>Fields</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>publication_id</td>
<td>int</td>
</tr>
<tr>
<td>publication_status</td>
<td>string</td>
</tr>
<tr>
<td>publication_title</td>
<td>string</td>
</tr>
<tr>
<td>publication_authors</td>
<td>string array</td>
</tr>
<tr>
<td>publication_catagory</td>
<td>string array</td>
</tr>
<tr>
<td>publication_keywords</td>
<td>string array</td>
</tr>
<tr>
<td>conference_name</td>
<td>string</td>
</tr>
<tr>
<td>publisher_name</td>
<td>string</td>
</tr>
<tr>
<td>publication_volume</td>
<td>string</td>
</tr>
<tr>
<td>publication_page_number</td>
<td>string</td>
</tr>
<tr>
<td>publication_isbn</td>
<td>string</td>
</tr>
<tr>
<td>publication_abstract</td>
<td>string</td>
</tr>
<tr>
<td>publication_place</td>
<td>string</td>
</tr>
<tr>
<td>publication_date</td>
<td>string</td>
</tr>
<tr>
<td>publication_type</td>
<td>string</td>
</tr>
<tr>
<td>publication_year</td>
<td>int</td>
</tr>
<tr>
<td>publication_website</td>
<td>string</td>
</tr>
</tbody>
</table>
Key Steps

- Convert the features from the document
- Create a initial list of keywords using available data
- Extract the keywords - NC value Algorithm
- Create Document Model
- Cluster Documents
- Create Topic Model
- Label Clusters
- Configure Solr
- Setup Web Server
- Develop visual aids
Information seeking and discovery

- Visual interfaces are geared towards more exploratory activities in contrast to goal-oriented search.
- How can exploratory approaches be complementary to or extensions of search?
- Using search limit the collection used by visualization
- What is the place for search in exploration?
- To understand the underlining trends
Overview of the system

Documents
(text fields:title,abstract)
(name:document ID)

Information Analysis System

Potential Keyword List

NC Value Algorithm

Document Vectors
(name: documentID,Content:keywords)

Clustering(LDA modeling of corpus)

Topics
(topWords,topDocuments)

Inverse Index
(Keyword:document of occurrence)

Keyword List
(keyword,score)

Data Wrapper

TimeLine

Graph

Area

Database

<table>
<thead>
<tr>
<th>Title 1</th>
<th>Title 2</th>
<th>Title 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
</tr>
<tr>
<td>Value 4</td>
<td>Value 5</td>
<td>Value 6</td>
</tr>
<tr>
<td>Value 7</td>
<td>Value 8</td>
<td>Value 9</td>
</tr>
<tr>
<td>Value 10</td>
<td>Value 11</td>
<td>Value 12</td>
</tr>
</tbody>
</table>

Wikipedia
MOBIX: System for managing MOBILITY using Information eXchange

We propose a different approach to determining network availability of mobile nodes which leverages on the fact that nodes on the move will meet other nodes who will be able to show conditions of networks they have recently encountered. This paper presents MOBIX, a system where nodes exchange information about network conditions using short-range communication such as Bluetooth. Our simulation results show that the required number of nodes needed for 100% success is not unrealistic for densely populated metropolitan areas. Even with relatively low population densities, we can expect a data store hit more than 50% of the time. Although our evaluation used WiFi, our scheme can easily be extended for other technologies such as GSM and WiMax.

Conference Paper | ACM | Moblows PhD Forum | NULL | Conference Paper | 2000-08-29

On the Impact of Modelling Choices for Distributed Information Spread -- A Comparative Study

We consider a distributed shuffling algorithm for sharing data in a distributed network. Nodes executing the algorithm periodically contact each other, and exchange data. The behavior of the algorithm is probabilistic in nature; a node chooses a random peer, and sends a random subset of its local data. Moreover, the algorithm exhibits non-deterministic behavior; the order in which nodes initiate an exchange is not specified. For the shuffling algorithm we build several formal models using the probabilistic model checker PRISM. Despite the well known state-space explosion problem, we were able to model a network of up to 15 nodes. In addition, we implement two equational models in Matlab, a discrete model and a continuous alternative, as well as the algorithm itself in the peer-to-peer network simulator PeerSim. By comparing different modelling frameworks, we further explore the impact of modelling choices, such as different scheduling policies and the notion of rounds. The comparison of different models allowed us to discover hidden assumptions in these alternative modelling frameworks, which helps with the interpretation of the obtained results.


Using Architecture Integration Patterns to Composing Enterprise Mashups

Enterprise mashups deal with cooperate data and various sources of information to compose new value-added applications. The architecture design of enterprise mashups encompasses integration issues - it needs to integrate heterogeneous data and/or compose new situational applications from existing infrastructures. We envisage that architecture integration patterns can be applied not only as architecture solutions to mashup development, but also to help us develop practical mashup techniques. In this paper, we combine several common architecture integration patterns, namely Pipes and Filters, Data Federation, and Model-View-Control to compose enterprise mashups. A number of techniques are also developed to customize these patterns for the specific mashup needs. We illustrate our approach by a property valuation service derived from the real-world setting.

Information seeking and discovery

- Vantage points: an entire collection and the individual artefact.
- Place for search in exploration: user may explorer individual categories
- Exploratory approach extends of search
Evaluation

- Effectiveness of visualization was evaluated by previous study.
- Clustering was evaluated based on results obtained by using 20 category Reuters data set.
Conclusion

- Provides insight to the research trends over the year.
- Topic labelling can be further improved by using larger and diverse data set.
- Seeking even intrusting patterns.
Question/ Answers