Search-Based Applications

At the Confluence of Search and Database Technologies

Synthesis Lectures on Information Concepts, Retrieval, and Services

Editor
Gari Marchionini, *University of North Carolina, Chapel Hill*

Synthesis Lectures on Information Concepts, Retrieval, and Services is edited by Gary Marchionini of the University of North Carolina. The series will publish 50- to 100-page publications on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. The scope will largely follow the purview of premier information and computer science conferences, such as ASIST, ACM SIGIR, ACM/IEEE JCDL, and ACM CIKM. Potential topics include, but not are limited to: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Search-Based Applications - At the Confluence of Search and Database Technologies
Gregory Grefenstette and Laura Wilber
2010

Information Concepts: From Books to Cyberspace Identities
Gary Marchionini
2010

Estimating the Query Difficulty for Information Retrieval
David Carmel and Elad Yom-Tov
2010

iRODS Primer: Integrated Rule-Oriented Data System
Arcot Rajasekar, Reagan Moore, Chien-Yi Hou, Christopher A. Lee, Richard Marciano, Antoine de Torcy, Michael Wan, Wayne Schroeder, Sheau-Yen Chen, Lucas Gilbert, Paul Tooby, and Bing Zhu
2010

Collaborative Web Search: Who, What, Where, When, and Why
Meredith Ringel Morris and Jaime Teevan
2009

Multimedia Information Retrieval
Stefan Rï¿½ger
2009

Online Multiplayer Games
William Sims Bainbridge
2009

Information Architecture: The Design and Integration of Information Spaces
Wei Ding and Xia Lin
2009

Reading and Writing the Electronic Book
Catherine C. Marshall
2009

Hypermedia Genes: An Evolutionary Perspective on Concepts, Models, and Architectures
Nuno M. Guimarï¿½es and Luï¿½s M. Carrico
2009

Understanding User-Web Interactions via Web Analytics
Bernard J. (Jim) Jansen

2009

XML Retrieval
Mounia Lalmas
2009

Faceted Search
Daniel Tunkelang
2009

Introduction to Webometrics: Quantitative Web Research for the Social Sciences
Michael Thelwall
2009

Exploratory Search: Beyond the Query-Response Paradigm
Ryen W. White and Resa A. Roth
2009

New Concepts in Digital Reference
R. David Lankes
2009

Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation
Michael G. Christel
2009

Search-Based Applications - At the Confluence of Search and Database Technologies

Gregory Grefenstette and Laura Wilber

ISBN: 978-3-031-01146-7 paperback
ISBN: 978-3-031-02274-6 ebook

DOI 10.1007/ 978-3-031-02274-6

A Publication in the Springer series
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES

Lecture #17
Series Editor: Gari Marchionini, *University of North Carolina, Chapel Hill*
Series ISSN
Synthesis Lectures on Information Concepts, Retrieval, and Services
Print 1947-945X Electronic 1947-9468

Search-Based Applications

At the Confluence of Search and Database Technologies

Gregory Grefenstette and Laura Wilber
Exalead, S.A.

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #17

ABSTRACT

We are poised at a major turning point in the history of information management via computers. Recent evolutions in computing, communications, and commerce are fundamentally reshaping the ways in which we humans interact with information, and generating enormous volumes of electronic data along the way. As a result of these forces, what will data management technologies, and their supporting software and system architectures, look like in ten years? It is difficult to say, but we can see the future taking shape now in a new generation of information access platforms that combine strategies and structures of two familiar – and previously quite distinct – technologies, search engines and databases, and in a new model for software applications, the Search-Based Application (SBA), which offers a pragmatic way to solve both well-known and emerging information management challenges as of now. Search engines are the world's most familiar and widely deployed information access tool, used by hundreds of millions of people every day to locate information on the Web, but few are aware they can now also be used to provide precise, multidimensional information access and analysis that is hard to distinguish from current database applications, yet endowed with the usability and massive scalability of Web search. In this book, we hope to introduce Search Based Applications to a wider audience, using real case studies to show how this flexible technology can be used to intelligently aggregate large volumes of unstructured data (like Web pages) and structured data (like database content), and to make that data available in a highly contextual, quasi real-time manner to a wide base of users for a varied range of purposes. We also hope to shed light on the general convergences underway in search and database disciplines, convergences that make SBAs possible, and which serve as harbingers of information management paradigms and technologies to come.

KEYWORDS

search-based applications, search engines, semantic technologies, natural language processing, human-computer information retrieval, data retrieval, online analytical processing, OLAP, data integration, alternative data access platforms, unified information access, NoSQL, mash-up technologies

Contents

Acknowledgments

Glossary

1 Search Based Applications

1.1 Introduction
1.1.1 What is a Search Based Application?
1.2 High Impact, Low Risk Solution for Businesses
1.3 Fertile Ground for Interdisciplinary Research
1.4 A Valuable Tool for Database Administrators
1.5 New Opportunities for Search Specialists
1.6 New Flexibility for Software Developers
1.6.1 Lecture Roadmap

2 Evolving Business Information Access Needs

2.1 Changing Times
2.2 The Need for High Performance and Scalability
2.3 The Need for Unified Access to Global Information
2.4 The Need for Simple Yet Secure Access

3 Origins and Histories

3.1 Search Engines
3.2 Databases
3.3 What has Changed Recently
3.3.1 Search Engines Enter the Enterprise
3.3.2 Databases Go Online
3.3.3 Structural and Conceptual Changes

4 Data Models & Storage

4.1 Search Engines

4.1.1 Conceptual Data Model
4.1.2 Data Storage
4.1.3 Storage Framework
4.2 Databases
4.2.1 Conceptual Data Model
4.2.2 Data Storage
4.2.3 Storage Framework
4.3 What has Changed Recently
4.3.1 Search Engines
4.3.2 Databases

5 Data Collection/Population

5.1 Search Engines
5.1.1 Collection
5.1.2 Updating
5.2 Databases
5.2.1 Creation/Collection
5.2.2 Updating
5.3 What has Changed
5.3.1 Search Engines
5.3.2 Databases

6 Data Processing

6.1 Search Engines
6.1.1 Natural Language Processing
6.1.2 Relevancy Criteria
6.2 Databases
6.3 What has Changed
6.3.1 Search Engines
6.3.2 Databases

7 Data Retrieval

7.1 Search Engines
7.1.1 Querying
7.1.2 Output

7.2 Databases
7.2.1 Querying
7.2.2 Output
7.3 What's Changed?
7.3.1 Search Engines
7.3.2 Databases

8 Data Security, Usability, Performance, Cost

8.1 Search Engines
8.2 Databases
8.3 What has Changed
8.3.1 Search Engines

9 Summary Evolutions and Convergences

9.1 SBA-Enabling Search Engine Evolutions
9.1.1 Data Model
9.1.2 Data Storage
9.1.3 Data Collection
9.1.4 Data Processing
9.1.5 Data Retrieval & Output
9.1.6 Data Security, Usability, Performance, Cost
9.2 Convergence

10 SBA Platforms

10.1 What is an SBA Platform?
10.2 Information Access Platforms
10.3 SBA Platforms: Market Leaders
10.4 SBA Platforms: Other Vendors
10.5 SBA Vendors: COTS Applications

11 SBA Uses & Preconditions

11.1 When Are SBAs Used?
11.2 How Are SBAs Used?

12 Anatomy of a Search Based Application

12.1 SBAs for Structured Data
12.1.1 Data Collection
12.1.2 Data Processing
12.1.3 Data Updates
12.1.4 Data Retrieval & Analysis
12.2 SBAs for Unstructured Content
12.2.1 Data Collection
12.2.2 Data Processing
12.2.3 Data Updates
12.2.4 Data Retrieval & Analysis
12.3 SBAs for Hybrid Content

13 Case Study: GEFCO

13.1 Background
13.2 A Track & Trace Solution
13.3 Existing Drawbacks
13.4 Opting for a Search Based Application
13.5 First prototypes
13.6 Deployment
13.7 Future

14 Case Study: Urbanizer

14.1 Background
14.2 The Urbanizer Solution
14.3 How Urbanizer Works
14.4 What's Next

15 Case Study: National Postal Agency

15.1 Customer Service SBA
15.1.1 Background
15.1.2 Deployment
15.2 Operational Business Intelligence (OBI) SBA
15.2.1 Background
15.2.2 Deployment
15.3 Sales Information SBA for Telemarketing

15.3.1 Background
15.3.2 Deployment

16 Future Directions

16.1 The Influence of the Deep Web
16.1.1 Surfacing Structured Data
16.1.2 Opening Access to Multimedia Content
16.2 The Influence of the Semantic Web
16.3 The Influence of the Mobile Web
16.3.1 Mission-Based IR
16.3.2 Innovation in Visualization
16.4 And Continuing Database/Search Convergence

Bibliography

Authors' Biographies

Index

Acknowledgments

We would like to thank Gary Marchionini and Diane Cerra for inviting us to participate in this timely and important lecture series, with a special thank you to Diane for her assistance and patience in guiding us through the publication process. We would also like to thank Morgan & Claypool's reviewers, including Susan Feldman, Stephen Arnold and John Tait, for their thoughtful suggestions and comments on our manuscript. Ms. Feldman and Mr. Arnold are constant sources of insight for all of us working in search and information access-related disciplines, and we welcome Mr. Tait's remarks based on his long IR research experience at the University of Sunderland and his more recent efforts at advancing research in IR for patents and other large scale collections at the Information Retrieval Facility.

In addition, we are grateful to our colleagues and managers at Exalead for allowing us time to work on this lecture, and for providing valuable feedback on our draft manuscript, especially Olivier Astier, Stéphane Donzé and David Thoumas. We would also like to thank our partners and customers. They are the source of the examples provided in this book, and they have played a pioneering role in expanding the boundaries of applied search technologies, in general, and search-based applications, in particular.

Finally, we would like to thank our families. Their love sustains us in all we do, and we dedicate this book to them.

Gregory Grefenstette and Laura Wilber
December 2010

Glossary

Glossary

ACID	Constraints on a database for achieving Atomicity, Consistency, Isolation and Durability
Agility	The ease with which a computer application can be altered, improved, or extended
API	Application Programming Interface, specifies how to call a computer program, what arguments to use, and what you can expect as output
Application layer	Part of the Open System Interconnection model, in which an application interacts with a human user, or another application
Atomicity	The idea that a database transaction either succeeds or fails in its entirety
Availability	The percentage of time that data can be read or used.
Batch	A computer task that is programmed to run at a certain time (usually at night) with no human intervention
B2C	Business to Customer; B2C websites offer goods or services directly to users
B+ tree	A block-oriented data structure for efficient insertion

	and removal of data nodes
BI	Business Intelligence, views on data that aid users with business planning and decision making
BigTable	An internal data storage system used by Google, handles multidimensional key-value pairs
BSON	Binary JSON
Business application	Any information processing application used in running a business
Cache	A rapid computer memory where frequently or recently used data is temporarily stored
CAP theorem	One cannot achieve Consistency, Availability, and Partition tolerance at the same time
Category	A flat or hierarchic semantic dimension added to a document, or part of a document
Categorization	Assigning, usually through statistical means, one or more categories to text
CDM	Customer Data Management
Cloud services	Computer applications that are executed on computers outside the enterprise rather than in-house. Examples are SalesForce, Google Apps, Yahoo mail, etc.
Clustering	Grouping documents according to content similarity
CMS	Content Management System

Consistency	A quality of an information system in which only valid data is recorded; that is, there are not two conflicting versions of the same data
Connector	A program that extracts information from a certain file format, or from a database
Consolidation	Making all the data concerning one entity available in one output
COTS	Commercial off-the-shelf software
Crawl	Fetching web pages for indexing by following URLs found in each page
CRM	Customer Relationship Management, applications used by businesses to interact with customers
CSIS	Customer Service Information System
Data integration	Merging data from different data sources or different information systems
Data mart	A subset of data found in an enterprise information system, relevant for a specific group or purpose
Data warehouse	A database which is used to consolidate data from disparate sources
DBA	Database administrator, the person who is responsible for maintaining (and often designing) an organization' database(s)
Deep Web	Web pages that are dynamically generated as a result of form input and/or database querying

Directory	A listing of the files or websites in a particular storage system
DIS	Decision Intelligence System, a computer-based system for helping decision making
Document model	A model of seeing a database entity as a single persistent document, composed of typed fields and categories corresponding to the entity's attributes
Dublin Core Metadata	A standard for metadata associated with documents, such as Title, Creator, Publisher, etc.
Durability	A database quality that means that successfully completed transactions must persist (or be recoverable) in the case of a system failure
EDI	Electronic Data Interchange, an early database communication system
ETL	Extract-Transform-Load, any method for extracting all or part of a database and storing it in another database
Enterprise Search	Searching access-controlled, structured and unstructured data found within the enterprise
ERP	Enterprise Resource Planning
Evolutive Data Model	Model that can be easily extended with new fields or data types without rebuilding the entire data structure
Facet	A dimension of meaning that can be used for restricting search, for example *shirts* and *coats* are two facets that could be found on a shopping site

Field	A labeled part of a document in a search engine. Fields can be typed to contain text, numbers, dates, GPS coordinates, or categories
Firewall	A computer-implemented protection that isolates internal company data from outside access
File server	A service that provides sequential or direct access to computer files
Full-text engine	A system for searching any of the words found in documents, rather than just a set of manually assigned keywords
Garbage collection	A process for recovering memory, usually by recognizing deleted or out-of-date data
Gartner	An information technology research and advisory firm that reports on technology issues
GPS	Global Positioning System, a system of satellites for geolocating a point on the globe
Hash table	Hashing converts a data item into a single number, and the hash table maps this number to a list of items
Heuristics	Methods based more on demonstrated performance than theory, weighting words by their inverse frequency in a collection is an example
HTTP	HyperText Transfer Protocol, an application layer protocol for accessing web pages
IDC	International Data Corporation, a global provider of market intelligence and analysis concerning

	information technology
ILM	Information Lifecycle Management
IMAP	Internet Message Access Protocol, a format for transmitting emails
Index, inverted	A data structure that contains lists of words with pointers to where the words are found in documents
Index slice	One section of an inverted index which can be distributed over many different computer stores
Intranet	A secure network that gives authorized users Web-style access to an organization's information assets (e.g., internal documents and web pages)
IR	Information Retrieval, the study of how to index and retrieve information, usually from unstructured text
IS	Information System, a generic term for any computer system for storing and retrieving information
Isolation	The database constraint specifying that data involved in a transaction are isolated from (inaccessible to) other transactions until the transaction is completed to avoid conflicts and overwrites
IT	Information Technology, a generic term covering all aspects of using computers to store and manipulate information
JDBC	Java Database Connectivity, a Java version of ODBC

Join	In a relational database, gathering together data contained in different tables
JSON	JavaScript Object Notation, a standard for exchanging data between systems
Key-value store	A data storage and retrieval system in which a key (identifying an entity) is linked to the one or more values associated with that entity. This allows rapid lookup of values associated with an entity, but does not allow joins on other fields
Mash-up	A software application that dynamically aggregates information from many different sources, or output from many processes, in a single screen
MDM	Master Data Management, a system of policies, processes and technologies designed to maintain the accuracy and consistency of essential data across many data silos
Metadata	Typed data associated with a document, for example, Author, Date, Category
Mobile Web	Web pages accessible through a mobile device such as a smartphone
MySQL	A popular open source relational database
Normalized relational schema	A model for a relational database that is designed to prevent redundancies that can cause anomalies when inserting, updating, and deleting data
NoSQL	Not Only SQL, an umbrella term for large scale data storage and retrieval systems that use structures and

	querying methodologies that are different from those of relational database systems
OBI	Operational Business Intelligence, data reporting and analysis that supports decision making concerning routine, day-to-day operations
OCR	Optical Character Recognition, a technology used for converting paper documents or text encapsulated in images into electronic text, usually with some noise caused by the conversion
ODBC	Open Database Connectivity, a middleware for enabling and managing exchanges between databases
Offloading	Extracting information from a database application and storing it in a search engine application
OLAP	Online Analytical Processing, tools for analyzing data in databases
OLTP	Online Transaction Processing
Ontology	A taxonomy with rules that can deduce links not necessarily present in the taxonomy
Partition tolerance	Means that a distributed database can still function if some of its nodes are no longer available
Performance	The measure of a computer application’s rapidity, throughput, availability, or resource utilization
PHP	PHP: Hypertext Preprocessor, a language for programming web pages

PLM	Product Lifecycle Management, systems which allow for the management of a product from design to retirement
Plug-and-play	Modules that can be used without any reprogramming, “out of the box”
POC	Proof of concept, an application that proves that something can be done, though it may not be optimized for performance
Portal	A web interface to a data source
Primary key	In a relational database, a value corresponding to a unique entity, that allows tables to be joined for a given entity
RDBMS	Relational database management system
Redundancy	Storing the same data in two different places in a data base, or information system. This can cause problems of consistency if one of the values is changed and not the other
Relational model	A model for databases in which data is represented as tables. Some values, called primary keys, link tables together
Relevancy	For a given query, a heuristically determined score of the supposed pertinence of a document to the query
REST	Representational State Transfer, protocol used in web services, in which no state is preserved, but in which every operation of reading or writing is self sufficient

RFID	Radio Frequency Identification, systems using embedded chips to transmit information
RSS	Really Simple Syndication, an XML format for transmitting frequently updated data
R tree	An efficient data structure for storing GPS-indexed points and finding all the points in a given radius around a point
RDF	Resource Description Framework, a format for representing data as sets of triples, used in semantic web representations
SBA	Search Based Applications, an information access or analysis application built on a search engine, rather than on a database.
SCM	Supply Chain Management
Scalability	The desirable quality of being able to treat larger and larger data sets without a decrease in performance, or rise in cost
Search engine	A computer program for indexing and searching in documents
Semantic Web	Collection of web pages that are annotated with machine readable descriptions of their content
Semi-structured data	Data found in places where the data type can be surmised, such as in explicitly labeled metadata, or in structured tables on web pages
SEO	Search engine optimization, strategies that help a web page owner to improve a site's ranking in

	common web search engines
SERP	Search engine results page, the output of a query to a search engine
Silo	An imagery-filled term for an isolated information system
SMART system	An early search engine developed by Gerald Salton at Cornell
SOAP	Simple Object Access Protocol, a format for transmitting data between services
Social media	Data uploaded by identified users, such as in YouTube, FaceBook, Flickr
SQL	Structured Query Language, commonly used language for manipulating relational databases
Structured data	Data organized according to an explicit schema and broken down into discrete units of meaning, with units represented using consistent data types and formats (databases, log files, spreadsheets)
SVM	Support vector machine, used in classification
Table	Part of a relational database, a body of related information. Each row of the table corresponds to one entity, and each column, to some attribute of this entity
Taxonomy	A hierarchically typed system of entities, such as mammals being part of animals being part of living beings

TCO	Total cost of ownership, how much an application costs when all implicit and explicit costs are factored in over time
Timestamp	A chronological value indicating when some data was created
Top-k	The *k* highest ranked responses in a database system that can rank answers to a query
Transaction	In databases, a sequence of actions that should be performed as an uninterruptable unit, for example, purchasing a seat on a flight
Unstructured data	Data that is not formally or consistently organized, such as textual data (email, reports, documents) and multimedia content
URL	Universal Resource Locator, the address of a web page
Usability	The desirable quality of being able to be used by a large population of users with little or no training
Vertical application	An application built for a specific domain, such as pharmaceuticals, finance, or manufacturing. A horizontal application could be used in a number of different domains.
XML	eXtended Markup Language, a standard for including metadata in a document
W3C	World Wide Web Consortium
WYSIWYG	What You See Is What You Get

YPG	Yellow Pages Group, Canada