Information Retrieval
Overview

many slides courtesy James Allan @umass Amherst
some slides courtesy ChengXiang Zhai @Urbana-Champaign
administrivia

- INSTRUCTOR
  Virgil Pavlu  vip@ccs.neu.edu

- GRADER: if we get any, he/she will have final authority on homework grading issues

- http://ccs.neu.edu/~jaa/ISU535.06X2
overview

- what is information retrieval?
- how does it work?
- a simple retrieval model
what is IR?

12. CIKM 2003 New Orleans, Louisiana, USA
www.informatik.uni-bielefeld.de/~dyb/db/conf/cikm/cikm2003.html - 29k - Cached - Similar pages

CIKM 2003 review
CIKM 2003 highlights. 12th ACM International Conference on Information and Knowledge Management, 3-8 November, New Orleans...
smu.colorado.edu/papers/cikm03_report.html - 22k - Cached - Similar pages

Collaborative Filtering Making List Archive: [english@simi] CFP
ACM CIKM 2003 Call For Papers, 12th International Conference on Information and Knowledge... call for papers submitted to CIKM 2003 will be accepted...
www.cs.cornell.edu/1896/0097.html - 17k - Cached - Similar pages

TOC
Proceedings of the twelfth international conference on Information and Knowledge Management, New Orleans, LA, USA, November 3 - 6, 2003...
portal.acm.org:acm.cfar?id=89655&type=proceeding - Similar pages

[Aashi][ CIKM 2003
mail.aoe.org/formal/news-2003/Sep/030124.html - 17k - Cached - Similar pages

proo CIKM 2003
File Format: PDF/Adobe Acrobat - View as HTML
Page 2. Outline of the talk: how to build a partition...
www.cs.umbc.edu/~nicholas/clusterig/pdf - Similar pages

Tutorial on Document Clustering
www.cs.umbc.edu/~nicholas/clusterig/pdf - 9k - Cached - Similar pages

Conference on Information and Knowledge Management (CIKM)
CIKM has a strong tradition of workshops devoted to emerging areas of database...
The CIKM 2004 web page, The CIKM 2003 Web Page, The CIKM 2002 Web Page...
www.cikm.org - 7k - Cached - Similar pages

CIKM 2003 New Orleans, USA, November 2003
Miniaturansicht
www.bibliothek.de/libertys, photo?folder=2003-11a_cikm03&picture=7 - 2k - Cached - Similar pages
Conference on Information and Knowledge Management (CIKM)
www.cikm.org
More pages from cikm.org

Department of Computer Science at LSU
News & Events, Department of Computer Science, 293 Coates Hall, Phone: (225) 578-1495
Visit us at: http://www.cs.lsu.edu
More pages from cs.lsu.edu

ACM CIKM 2003 Call For Papers
More pages from cs.wisc.edu

MMDB03
... conjunction with The 12th International Conference on Information and Knowledge Management (ACM CIKM 2002) Call for... www.cs.fl.edu/mmdb03
More pages from cs.fl.edu

CIKM 2002 Homepage
The CIKM 2002 Technical Program has been updated. Hotel information has been updated in the Local Arrangements information forum for presentation... The 12th International Conference on Information and Knowledge Management - CIKM 2002
More pages from cikm.org

DBWorld Message
Only the highest calibre papers submitted to CIKM 2003 will be accepted. We have a special interest in paper... For Papers 12th International Conference on Information and Knowledge Management (CIKM2003)... www.cs.wisc.edu/dbworld/messages/2003-06/10571827389.html
More pages from cs.wisc.edu

Past Events
... July 28 - August 1, 2003. SIGIR 2003 (Toronto) - 26th Annual International ACM SIGIR Conference... May 28, www.sigir.org/events/events-past.html
More pages from sigir.org

Upcoming Conference Deadlines
More pages from www-users.cs.umn.edu
Conference on Information and Knowledge Management (CIKM)

more info from http://www.cikm.org/ - 6 KB

11. CIKM 2003: New Orleans, Louisiana, USA
ISBN 1-58113-723-0
more info from http://www.informatik.uni-erlangen.de/ley/db/conf/cikm/cikm2003.html - 65 KB

Department of Computer Science at LSU
News & Events. Department of Computer Science. 288 Coates Hall. Phone: (225) 578-1485
more info from http://cit.coe.lsu.edu/ - 3 KB

CIKM
CIKM 2003 Home Page. 11. CIKM 2002: McLean, Virginia, USA.
more info from http://sunsite.informatik.wias-berlin.de/dbconf/CIKM - 10 KB

ACM WIDM 2003
Past WIDM's. Conference on Information and Knowledge Management (CIKM 2003). Hotel Inter-Continental, New
more info from http://www.cs.ubc.ca/~wilm/widm/ - 12 KB

ACM CIKM 2003 Call For Papers
ACM CIKM 2003 Call For Industry Track Presentations. We solicit high-quality Industry Track presentations which p
more info from http://www.cs.wisc.edu/dbworld/messages/2003-08/087179548.html - 35 KB

ACM CIKM 2003: Agents Portal: Agent and Multi-agent Technology Resources
... PRELIMINARY CALL FOR PAPERS. http://cikm.org/2003/Twelfth International Conference on Information and ...

MMBI 03
...in conjunction with the 12th International Conference on Information and Knowledge Management (ACM CIKM 200
more info from http://www.cs.fsu.edu/mmdb03 - 5 KB

CIKM 2002 Homepage
The CIKM 2002 Technical Program has been updated. Hotel information has been updated in the Local Arranger for pr
presentation. The 12th International Conference on Information and Knowledge Management (CIKM 2003)...
more info from http://cikm.org/2002 - 15 KB

DBWorld Message
Only the highest caliber papers submitted to CIKM 2003 will be accepted. We have a special interest in papers the In
International Conference on Information and Knowledge Management (CIKM)...
more info from http://www.cs.wisc.edu/dbworld/messages/2003-03/051827339.html - 4 KB
metasearch
text management applications

- Access
- Mining
- Create Knowledge
- Organization
- Add Structure/Annotations

Select information
text management applications

Retrieval Applications
- Summarization
- Filtering
- Search
- Categorization

Information Organization

Mining Applications
- Visualization
- Mining
- Extraction
- Clustering

Natural Language Content Analysis

Text

Information Access

Knowledge Acquisition
A dog is chasing a boy on the playground.

Semantic analysis
Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

Syntactic analysis
A dog (dog) is (is) chasing (chasing) a boy (boy) on (on) the playground (playground).

Lexical analysis (part-of-speech tagging)
Det Noun Aux Verb Det Noun Prep Det Noun
Noun Phrase Complex Verb Noun Phrase Prep Phrase Noun Phrase

Inference
Scared(x) if Chasing(_,x,__).
Scared(b1)

A person saying this may be reminding another person to get the dog back...

Pragmatic analysis (speech act)
what we can do in NLP

**A dog is chasing a boy on the playground**

- **POS Tagging:** 97%
- **Parsing:** partial >90% (?)
- **Speech act analysis:** ???
- **Semantics: some aspects**
  - Entity/relation extraction
  - Word sense disambiguation
  - Anaphora resolution
- **Inference:** ???

**Sentence:**

- **Det**
- **Noun**
- **Aux**
- **Verb**
- **Det**
- **Noun**
- **Prep**
- **Det**
- **Noun**
natural language processing

Arabic text

How can a computer make sense out of this string?

Morphology
- What are the basic units of meaning (words)?
- What is the meaning of each word?

Syntax
- How are words related with each other?

Semantics
- What is the “combined meaning” of words?

Pragmatics
- What is the “meta-meaning”? (speech act)

Discourse
- Handling a large chunk of text

Inference
- Making sense of everything
search (ad-hoc IR)

User

query

"robotics applications"

Retrieval System

relevant docs

Robotics

non-relevant docs

others

database/collection

text docs
information filtering

- Stable & long term interest, dynamic info source
- System must make a delivery decision immediately as a document “arrives”
collaborative filtering
Pre-given categories and labeled document examples (Categories may form hierarchy)
Classify new documents
A standard supervised learning problem
clustering

- Discover “natural structure”
- Group similar objects together
- Object can be document, term, passages
## IR vs Databases

<table>
<thead>
<tr>
<th></th>
<th>Databases</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td><strong>Structured</strong></td>
<td><strong>Unstructured</strong></td>
</tr>
<tr>
<td><strong>Fields</strong></td>
<td><strong>Clear semantics</strong> (SSN, age)</td>
<td><strong>No fields (other than text)</strong></td>
</tr>
<tr>
<td><strong>Queries</strong></td>
<td><strong>Defined</strong> (relational algebra, SQL)</td>
<td><strong>Free text (&quot;natural language&quot;), Boolean</strong></td>
</tr>
<tr>
<td><strong>Recoverability</strong></td>
<td><strong>Critical</strong> (concurrency control, recovery, atomic operations)</td>
<td><strong>Downplayed, though still an issue</strong></td>
</tr>
<tr>
<td><strong>Matching</strong></td>
<td><strong>Exact</strong> (results are always “correct”)</td>
<td><strong>Imprecise (need to measure effectiveness)</strong></td>
</tr>
</tbody>
</table>
related areas

- Models
- Algorithms
- Statistics
- Optimization
- Machine Learning
- Pattern Recognition
- Data Mining
- Natural Language Processing
- Applications
  - Web, Bioinformatics...
- Library & Info Science
- Databases
- Software engineering
- Computer systems
- Systems
publications/societies

Learning/Mining
- ICML
- ICML, NIPS, UAI
- ACM SIGKDD
- AAAI
- NLP
- ACL
- COLING, EMNLP, ANLP

Applications
- ISMB
- RECOMB, PSB
- WWW

Info Retrieval
- ACM SIGIR
- ACM CIKM, TREC
- JCDL

Info. Science
- ASIS

Databases
- ACM SIGMOD
- VLDB, PODS, ICDE

Software/systems

Statistics

??
SIGIR 2003 topics

- Formal Models, Language Models, Fusion/Combination
- Text Representation and Indexing, XML and Metadata
- Performance, Compression, Scalability, Architectures, Mobile Applications
- Web IR, Intranet/Enterprise Search, Citation and Link Analysis, Digital Libraries, Distributed IR
- Cross-language Retrieval, Multilingual Retrieval, Machine Translation for IR
- Video and Image Access, Audio and Speech Retrieval, Music Retrieval
- Machine Learning for IR, Text Data Mining, Clustering, Text Categorization
- Topic Detection and Tracking, Content-Based Filtering, Collaborative Filtering, Agents
- Summarization, Question Answering, Natural Language Processing for IR, Information Extraction, Lexical Acquisition
- Interfaces, Visualization, Interactive IR, User Models, User Studies
- Specialized Applications of IR, including Genomic IR, IR in Software Engineering, and IR for Chemical Structures
overview

- what is information retrieval?
- how does it work?
- a simple retrieval model
basic approaches

- boolean
- geometric: vector space model
- probabilistic: language models
- statistical: bayesian networks
- graph like: page rank
relevant items are similar

• Much of IR depends upon idea that
  **similar vocabulary → relevant to same queries**

• Usually look for documents matching query words

• “Similar” can be measured in many ways
  – String matching/comparison
  – Same vocabulary used
  – Probability that documents arise from same model
  – Same meaning of text
• An effective and popular approach
• Compares words without regard to order
• Consider reordering words in a headline

- **Random**: beating takes points falling another Dow 355
- **Alphabetical**: 355 another beating Dow falling points
- **“Interesting”**: Dow points beating falling 355 another

- **Actual**: Dow takes another beating, falling 355 points
what is this about?

16 × said
12 × fat
8 × new
5 × food oil percent reduce taste Tuesday
4 × amount change health Henstenburg make obesity
3 × acids consumer fatty polyunsaturated US
2 × amounts artery Beemer cholesterol clogging director
down eat estimates expert fast formula impact initiative
moderate plans restaurant saturated trans win
1 × ...

added addition adults advocate affect afternoon age
Americans Asia battling beef bet brand Britt Brook Browns
calorie center chain chemically ... crispy customers cut ...
vegetable weapon weeks Wendys Wootan worldwide years
York
McDonald's slims down spuds
Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA. But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.
text representation

• Text representation
  – what makes a “good” representation?
  – how is a representation generated from text?
  – what are retrievable objects and how are they organized?

• Representing information needs
  – what is an appropriate query language?
  – how can interactive query formulation and refinement be supported?

• Comparing representations
  – what is a “good” model of retrieval?
  – how is uncertainty represented?
hypertext
hypertext
overview

• what is information retrieval?
• how does it work?
  • a simple retrieval model
retrieval process
statistical language model

\[ D = \begin{cases} 
\text{One fish, two fish, red fish, blue fish.} \\
\text{Black fish, blue fish, old fish, new fish.} 
\end{cases} \]

\[ \text{len}(D) = 16 \]

\[ P(\text{fish}|D) = \frac{8}{16} = 0.5 \]
\[ P(\text{blue}|D) = \frac{2}{16} = 0.125 \]
\[ P(\text{one}|D) = \frac{1}{16} = 0.0625 \]
\[ P(\text{eggs}|D) = \frac{0}{16} = 0 \]

... 

A “topic”
statistical language model

- Document came from a topic
- Did query come from *this* document’s topic?

- For each document, find probability its topic could have generated the query

\[
P(Q|T_D) \approx P(Q|D) = P(q_1, \ldots, q_t|D) = \prod_{i=1}^{t} P(q_i|D)\]

Independence assumption (Naïve Bayes)
statistical language model

\[
D_1 = \begin{cases} 
\text{This one, I think, is called a Yink.} \\
\text{He likes to wink, he likes to drink.} 
\end{cases}
\]

\[
D_2 = \begin{cases} 
\text{He likes to drink, and drink, and drink.} \\
\text{The thing he likes to drink is ink.} 
\end{cases}
\]

\[
D_3 = \begin{cases} 
\text{The ink he likes to drink is pink.} \\
\text{He likes to wink and drink pink ink.} 
\end{cases}
\]

Query “drink”
- \(P(\text{drink}|D_1) = 1/16\)
- \(P(\text{drink}|D_2) = 4/16\)
- \(P(\text{drink}|D_3) = 2/16\)

Query “pink ink”
- \(P(Q|D_1) = 0 \cdot 0 = 0\)
- \(P(Q|D_2) = 0 \cdot 1/16 = 0\)
- \(P(Q|D_3) = 2/16 \cdot 2/16 = 0.016\)

Query “wink drink”
- \(P(Q|D_1) = 0.004\)
- \(P(Q|D_2) = 0\)
- \(P(Q|D_3) = 1/16 \cdot 2/16 = 0.008\)
Information Retrieval?
- Indexing, retrieving, and organizing text by probabilistic or statistical techniques that reflect semantics without actually understanding
- Search engines

Core idea
- Bag of words captures much of the “meaning”
- Objects that use vocabulary the same way are related

Statistical language model
- Documents used to estimate a topic model
- Query reflects a topic, too
- Documents of topics that are likely to produce the query are most likely to be relevant
in this course

- http://ccs.neu.edu/~jaa/ISU535.06X2/schedulen.html