

Text Technologies

Victor Lavrenko
vlavrenk@inf

Copyright © Victor Lavrenko, 2011

What you will learn

- How to build a search engine
 - which search results to rank at the top
 - how to do it fast and on a massive scale
- How to evaluate a search algorithm
 - is system A really better than system B
- How to work with text
 - two web-pages discuss the same topic?
 - handle misspellings, morphology, synonyms
 - build algorithms for languages you don't know

© Victor Lavrenko, 2011

Overview

- Information Retrieval
 - Two main issues in IR: speed and accuracy
 - Documents, queries, relevance
 - Bag-of-words trick
- Overview of Search System Architectures
- Other IR tasks

© Victor Lavrenko, 2011

Information Retrieval (IR)

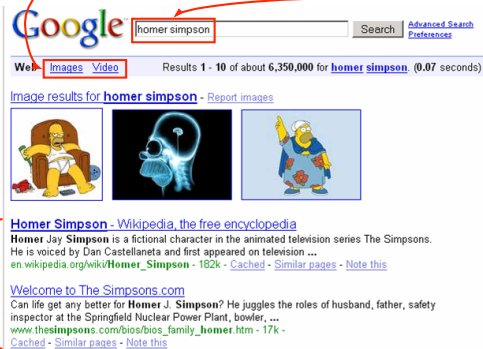
“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968)

- IR – core technology for text processing
- widely used in NLP/DBMS applications
- driving force behind web technologies

© Victor Lavrenko, 2011

IR in a nutshell

Find **documents** in response to the user's **query**



© Victor Lavrenko, 2011

Two main issues in IR

- Effectiveness
 - need to find **relevant** documents
 - needle in a haystack: Results 1 - 10 of about 6,350,000 for homer simpson.
 - very different from relational DBs (SQL)
- Efficiency
 - need to find them quickly: r homer simpson (0.07 seconds)
 - vast quantities of data (40b/120b pages)
 - thousands queries per second
 - data constantly changes, need to keep up
 - compared with other NLP areas IR is **very fast**

© Victor Lavrenko, 2011

Documents

- “documents” has a very wide meaning:
 - web-pages, emails, word/pdf/excel, news
 - photos, videos, musical pieces, code
 - answers to questions
 - product descriptions, advertisements
 - may be in a different language
 - may not have words at all (e.g. DNA)
- IR: match A against a large set of Bs
 - problem arises in many different domains

© Victor Lavrenko, 2011

Queries

- commercial engines:
 - query = a few keywords (“homer simpson”)
- query = expression of information need
 - describes what you want to find
 - can have many forms:
 - keywords, narrative, example “document”
 - question, photo, scribble, humming a tune
 - **#wsum**(0.9 **#field** (title, **#phrase** (homer,simpson))
0.7 **#and** (**#>** (pagerank,3), **#ow3** (homer,simpson))
0.4 **#passage** (homer, simpson, dan, castellaneta))



© Victor Lavrenko, 2011

Relevance

- at an abstract level, IR is about:
 - does item **D** **match** item **Q**? ...or...
 - is item **D** **relevant** to item **Q**?
- relevance a tricky notion
 - will the user like it / click on it?
 - will it help the user achieve a task?
 - is it novel (not redundant)?
- common take: *relevance* = *topicality* / *aboutness*
 - i.e. **D, Q** share similar “meaning”
 - about the same topic / subject / issue


© Victor Lavrenko, 2011

Why is *matching* a challenge?

- no clear semantics, contrast:
 - **author** = X123456 (“*Shakespeare, William*”) vs.
 - “play, frequently attributed to Shakespeare, is in fact”
- inherent ambiguity of language:
 - synonymy: “banking crisis” = “financial meltdown”
 - polysemy: “Homer” can be “Simpson” or “Greek”
- relevance highly subjective
 - Anomalous State of Knowledge (Belkin)
- relevance not observable (when we need it)
- on the web: counter SEOs / spam

© Victor Lavrenko, 2011

How do search engines do it?

- not with relational DBs
 - ok in niche domains (libraries)
 - “tagging” works for multi-media  homer, simpson, cartoon
 - spammers, loses “clarity” with scale
 - “semantic web” → inconsistent ontologies
- not by “understanding” the language
 - NLP brittle in unrestricted domains
 - good w. fixed structure/vocabulary (e.g. takeovers)
 - computationally expensive

© Victor Lavrenko, 2011

Relevant Items are Similar

- Key idea:
 - use similar vocabulary → similar meaning
 - similar documents relevant to same queries
- Similarity
 - string match
 - word overlap
 - P (same model)
 - ...

[How single stars lost their companions](#)

Space Daily - Sep 15, 2011

by Staff Writers Not all stars are loners. In our home galaxy, the Milky Way, about half of all stars have a companion and travel through space in a binary system. But explaining why some stars are in double or even triple systems while others are ...

[Coupled stars break up for the single life](#)

Astronomy Now Online - Gemma Lavender - Sep 16, 2011

Why some stars prefer to be single, while others are either paired up or in trios, could have been answered by a team of astronomers at the Max-Planck-Institute for Radio astronomy and the University of Bonn with the help of sophisticated computer ...

© Victor Lavrenko, 2011

Bag-of-words trick

- Can you guess what this is about:
 - beating falls 355 Dow another takes points
Dow takes another beating, falls 355 points
 - said fat fries McDonalds French obesity
does “French” refer to “France” here? why?
- Re-ordering doesn’t destroy meaning
 - individual words – “building blocks”
 - “bag” of words: a “composition” of “meanings”

© Victor Lavrenko, 2011

Bag-of-words trick (2)

- Most search engines use BOW
 - treat documents, queries as *bags* of words
 - a “bag” is a set with repetitions (multi-set, urn)
 - match = “degree of overlap” between *D, Q*
- Retrieval models
 - statistical models that use words as features
 - decide which docs most likely to be relevant
 - what should be the top 10 for “homer simpson”?
 - BOW makes these models tractable

© Victor Lavrenko, 2011

Bag-of-words: criticisms

- word meaning lost without context
 - true, but BOW doesn’t really discard context
 - it discards surface form / well-formedness of text
- what about negations, etc.?
 - “not, but he loves me” vs. “but he loves me not”
 - still discusses the same subject (him,me,love)
 - propagate negations to words: “but he not_loves me”
- does not work for all languages
 - no natural “word” unit Chinese, images, music
 - circumvent by “segmentation” or “feature induction”
 - break/aggregate until units reflect “aboutness”

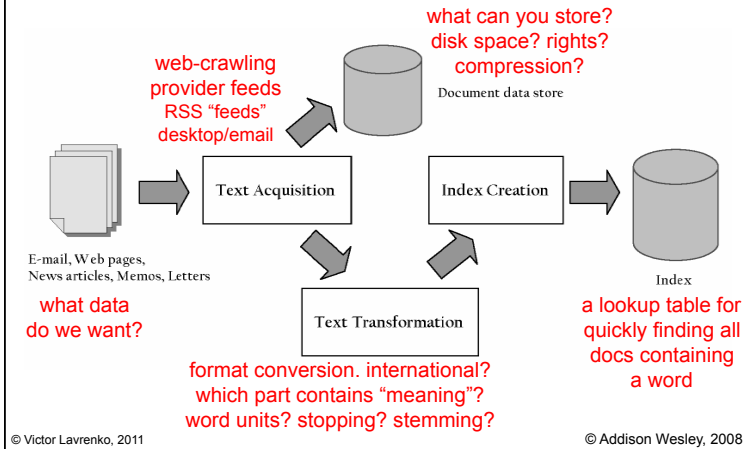
© Victor Lavrenko, 2011

Systems perspective on IR

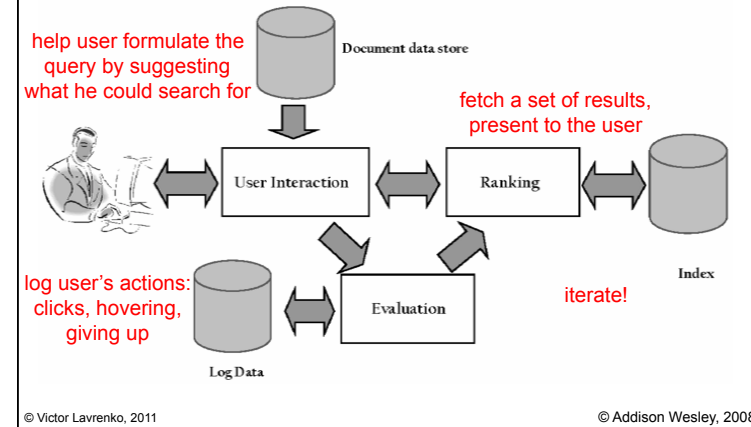
- get the data into the system
 - acquire the data from crawling, feeds, etc.
 - store the originals (if needed)
 - transform to BOW and “index”
- satisfy users’ requests
 - assist user in formulating query
 - retrieve a set of results
 - help user browse / re-formulate
 - log user’s actions, adjust retrieval model

© Victor Lavrenko, 2011

Indexing Process



Search Process



Other tasks

- some IR tasks don't fit "retrieval" view:
 - clustering: group similar items
 - discover "topics" in a scientific collection
 - classification: assign labels to items
 - predict if a news story will affect market
 - recommendation / collaborative filtering
 - guess which book the user might want to buy
 - event detection and tracking
 - detect novel events (e.g. new hurricane) in news

© Victor Lavrenko, 2011

Summary

- Information Retrieval (IR): core technology
 - selling point: IR is very fast, provides context
- Main issues: effectiveness and efficiency
- Documents, queries, relevance
- Bag-of-words trick
- Search system architecture:
 - indexing: get data into the system
 - searching: help users find relevant data

© Victor Lavrenko, 2011

Administrivia

© Victor Lavrenko, 2011

Course Information

- Lectures: [Mon/Thu 12-1pm](#) in OC.183 / HRB.LT
- Coursework due in weeks 4,6,8,10
- Practical labs in weeks 2,4,6,8
 - TA: Philipp Petrenz, [groups/rooms/times](#) TBD
- Textbook:
 - “[Search Engines: Information Retrieval in Practice](#)”
- Syllabus / assignments / notes:
 - www.inf.ed.ac.uk/teaching/courses/tts
- Ask questions on the forum:
 - www.forums.ed.ac.uk/viewforum.php?f=821
- Auditing: register as “observer” with ITO

© Victor Lavrenko, 2011

Assessment

- **(70%)** written exam: April 2012
- **(30%)** four programming assignments
 - due 4pm Monday in weeks 4,6,8,10
 - must use Python
 - “getting started” tutorial in this week
 - last year’s topics (subject to change)
 - intelligent web crawler
 - plagiarism detector
 - image search engine
 - PageRank on emails
- **all** deadline extensions through ITO

© Victor Lavrenko, 2011

Ethics

- Submitted work must be your own
 - ok to discuss assignments with each other
 - not ok to share code / data / figures
 - suggestion: talk, don’t write anything down
- Always cite your sources
 - including web, fellow students, other courses
 - exception: lectures from this course
 - when in doubt: cite

© Victor Lavrenko, 2011

c.o. James Allan