Semantics of Search Technologies

Dr. Kavi Mahesh

Director, Indian Institute of Information Technology - Dharwad

Library Technology Conclave

Goa University, 24 January 2018

Search Problem

- Search engines: *I shall match keywords and*
  my user shall be happy with the results

- Many users: *this is how search engines work*

- Not in scholarly search!

The Real Search Problem

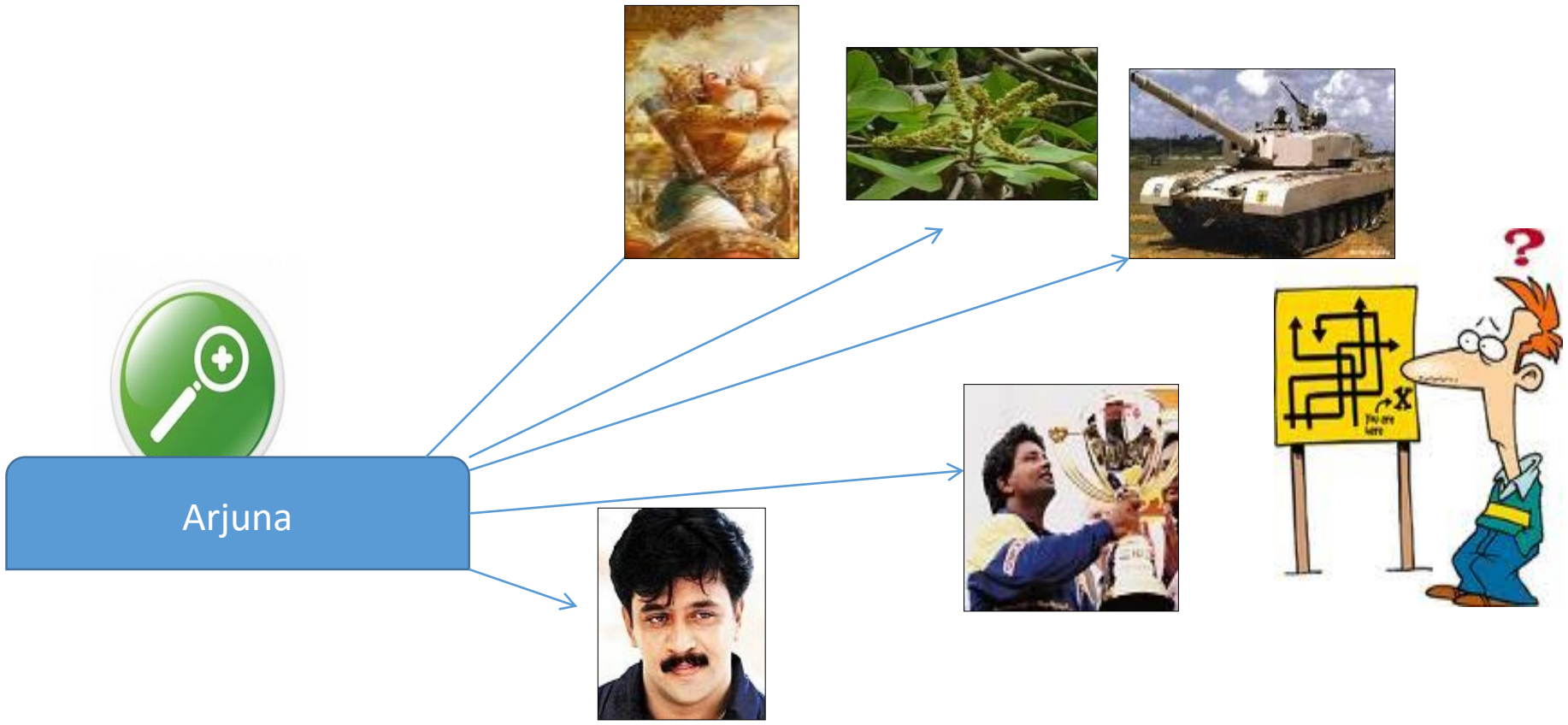
- User has a need for information or knowledge
- User has an imperfect way of conveying that need
  - sometimes just wants to explore
- The search engine has to retrieve:
  - Only relevant content from the repository
  - All of the relevant content in the repository
  - In ranked order of decreasing relevance
  - Who says what is relevant? *The user*

  (also extract information, aggregate, synthesize and answer questions….)

The Problem of Relevance: Precision/Recall

- “I promise to find relevant information, all the relevant information and only relevant information.”
- Precision: What percentage of results are actually relevant
- Recall: What percentage of relevant content was retrieved
- Can’t achieve both perfect precision and perfect recall!

Example: Arjuna



Arjuna

Search

About 34,100,000 results

Advanced search

Everything
Images
Videos
News
Shopping
More

Any time
Past hour
Past 24 hours
Past 2 days
Past week
Past month
Past year

More search tools

Arjuna - Wikipedia, the free encyclopedia
Arjuna in Indian mythology is the greatest warrior on earth and is one of the Pandavas, the heroes of the Hindu epic Mahābhārata. **Arjuna**, whose name. means ...
en.wikipedia.org/wiki/**Arjuna** - Cached - Similar

Arjuna Wholefoods
Arjuna Wholefoods Cambridge have a range of vegan and vegetarian wholefoods to buy online.
www.**arjuna**wholefoods.co.uk/ - Cached - Similar

Home | **Arjuna**
Commercial provider of distributed transaction processing technology, including CORBA Object Transaction Service compliant systems implemented in C++ and ...
www.**arjuna**.com/ - Cached - Similar

Arjuna bark extract supplement, benefit and side effects by Ray ...
Arjuna supplement health benefit and side effects, dosage and research studies by Ray Sahelian, M.D. 500 mg tablets. Ancient medical scientists have ...
www.raysahelian.com/**arjuna**.html - Cached - Similar

Arjuna
The Living Essence Foundation - details of schedules; transcripts of dialogues; books and tapes.
www.livingessence.com/ - Cached - Similar

Arjuna (TV) - Anime News Network

Arjuna - A Deusa do Tempo (Portuguese). **Arjuna** - la **...** Chikyū Shōjo **Arjuna** (Japanese). Earth Girl **...** Animax South Africa's Official Earth Girl **Arjuna** Website **...**
www.animenewsnetwork.com/encyclopedia/anime.php?id=602 - Cached - Similar

Arjuna - Newcastle University

These pages give information about the teaching and research activities of the School of Computing Science at Newcastle University, Newcastle upon Tyne, **...**
arjuna.ncl.ac.uk/ - Cached - Similar

Arjuna Productions :: Rich Media Solutions - Video on Internet **...**

Arjuna productions specialises in innovative and content driven audiovisual, multimedia and Internet communication packages that express what YOU want to **...**
www.**arjuna**.be/ - Cached - Similar

Arjuna - Free WordPress Theme - SRS Solutions

Arjuna is a regularly updated free WordPress theme. **Arjuna** stands for elegance, accessibility, and attention to detail.
www.srssolutions.com/en/downloads/**arjuna**_wordpress_theme - Cached - Similar

arjuna del toso

In my way back home from the office sometimes I have to pass this traffic light, it's at one end of Butt Bridge at the corner with Eden Quay, in Dublin (Ireland).
arjuna.deltoso.net/ - Cached - Similar

Shopping results for **Arjuna**

Himalaya

Arjuna

Full Spectrum

Arjuna 550mg

Solaray -

More than Keyword Matching ?

- Language is a natural phenomenon
- It is not “neat”
- We don’t know how we process language
- Example: How do we read?

If yuo cna raed tihs, yuo hvae a sgtrane mnid too

Cna yuo raed tihs?

I cdnuolt blveiee taht I cluod aulaclty uesdnatnrd

waht I was rdanieg.

The Phaonmneal pweor of teh hmuan mnid, aoccdrnig to a rscheearch at Cmabridge Uinervisy, it dseno't

mtaetr ni waht oerdr teh ltteres in a wrod are, teh olny iproamtnt tihng is taht the frsit and isat ltteer be in teh rghit

pclae. Teh rset can be a taotl mses and you can sitll raed it

whotuit a pboerlm. Tihs is bcuseae teh huamn mnid deos not raed ervey lteter by istlef, but teh wrod as a wlohe.

Azanmig huh? Yaeh, and I awlyas tghuhot slpeling was ipmorantt!

Back to the problem of search…

More than Keyword Matching ?

- Page Rank algorithm
  - *More than what is in the document, let me go by who is recommending it*
  - Especially useful in citation networks
- Natural language semantics
- Data-driven or statistical modelling
  - Language models
  - User models
  - User behaviour models
- Thesaurus, ontology
- Collaborative filtering and recommendation algorithms
  - *You are not unique!*
  - *Tell me who your friends are and I will tell you what you should read!*
- Popularity? Altmetrics?

Increasing Recall: Term Expansion

- Stemming (or morphological analysis)
  - To what extent?
- Thesaurus expansion
  - E.g., WordNet syn-sets
- The problem with synonyms
  - Desk = Table
  - Table = Chart

Beyond Morphology and Thesauri

- Can we represent the semantics (or meaning) of each word?
  - Use a dictionary?
- Can we extract or synthesize the meaning of a document from the meanings of its words?
  - Is Google doing it?
  - Is anybody doing it?

The Real True Semantics

- Ontology: how is each term different from every other term?
- Example: can we build an ontology for “movement in water”:
  - Swim
  - Float
  - Dive
  - Sink
  - Drown
  - Walk in water
  - Swim under water
  - Sea surfing
  - Rise to the surface
  - Jump out of water

Underlying Problems

- Ambiguity and polysemy
  - E.g., server
- Context
- Ellipsis
- Phrasal semantics
  - E.g., Non-Banking Finance Corporation
  - E.g., There is no alternative
  - E.g., The new world atlas of artificial night sky brightness
- Non-literal usage
  - E.g., *Don’t throw out the baby with the metrics bathwater*
- Translation

What are they good for….

- Page Rank algorithm
  - Ranking of papers and authors
  - Journals, publishers and universities?
- Data-driven or statistical modelling
  - Too much data needed, too open ended
- Natural language semantics
  - Not accurate enough at present
- Thesaurus, ontology
  - Who will build them?
- Collaborative filtering and recommendation algorithms
  - If user community is known

What can we do practically?

A real need

- High-quality open, publication and citation database
  - Publish a Linked Open Dataset (LOD) in RDF format?
  - LOC-DB?
  - I4OC?
- Why pay for proprietary services built on incomplete databases?

- Can the Indian LIS community make this happen?
  - Make in India?

Thank you!