

COM6115: Text Processing

Information Retrieval: Task definition

Mark Hepple

Department of Computer Science
University of Sheffield

Overview

- Definition of the information retrieval problem
- Approaches to document indexing
 - ◊ manual approaches
 - ◊ automatic approaches
- Automated retrieval models
 - ◊ boolean model
 - ◊ ranked retrieval methods (e.g. vector space model)
- Term manipulation:
 - ◊ stemming, stopwords, term weighting
- Web Search Ranking
- Evaluation

Google search

jaguar

[Jaguar International - Market selector page](#)

www.jaguar.com/

Official worldwide web site of **Jaguar** Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

[Jaguar International - Home](#)



www.jaguar.com/gi/en/

8 Jul 2009

Our mission at **Jaguar** has been to create and build beautiful fast cars. The XK, XF, and XJ bring the ...

[More videos for jaguar »](#)

[Jaguar Cars - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Jaguar_Cars

Jaguar Cars Ltd, known simply as **Jaguar** is a British luxury and sports car manufacturer, headquartered in Whitley, Coventry, England. It is part of the **Jaguar** ...

[Jaguar - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Jaguar

The **jaguar** is a big cat, a feline in the *Panthera* genus, and is the only *Panthera* species found in the Americas. The **jaguar** is the third-largest feline after the tiger ...

Google search (contd)

jaguar south america

[Jaguar - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Jaguar

The **jaguar's** present range extends from Southern United States and Mexico across much of Central **America** and **south** to Paraguay and northern Argentina.

[Jaguar Cars](#) - [Jaguar \(disambiguation\)](#) - [Jacksonville Jaguars](#) - [Jaguarundi](#)

[Images for jaguar south america](#) - Report images



[South America - Jaguar](#)

library.thinkquest.org/5053/SouthAmerica/jaguar.html

Jaguars are magnificent cats that prowl the **South American** jungles. They are fascinating to learn about! To jump to a section, use our Quick Jump below by ...

[Jaguars, Jaguar Pictures, Jaguar Facts - National Geographic](#)

animals.nationalgeographic.com/animals/mammals/jaguar/

Jaguars are the largest of **South America's** big cats. They once roamed from the southern tip of that continent north to the region surrounding the U.S.-Mexico ...

Google search (contd)

black fast jaguar

[Jaguars, Jaguar Pictures, Jaguar Facts - National Geographic](#)

animals.nationalgeographic.com/animals/mammals/jaguar/

Learn all you wanted to know about **jaguars** with pictures, videos, photos, facts, ... **Fast** Facts. Type: Mammal; Diet: Carnivore; Average life span in the wild: 12 to ... Most **jaguars** are tan or orange with distinctive **black** spots, dubbed "rosettes" ...

[Jaguar XKR black fast on trackday - YouTube](#)



www.youtube.com/watch?v...

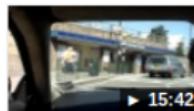
2 Jul 2012 - 16 sec - Uploaded by PrestigeCarCompany

Enjoy this **Jaguar** XKR trackday video.

<http://www.prestigecarcompany.co.uk/> Prestige Super Car Sales

...

[Jaguar XKR Black Pack fast racing - YouTube](#)



www.youtube.com/watch?v...k

16 Aug 2011 - 16 min - Uploaded by MrBobkumar

My XKR being driven hard thru town, hear the sounds of this beast....left window open so noise from air etc....but ...

[Jaguar XKR black fast on trackday. Rear Shot - YouTube](#)



www.youtube.com/watch?v=RmvW...

2 Jul 2012 - 17 sec - Uploaded by PrestigeCarCompany

Enjoy this **Jaguar** XKR trackday video.

<http://www.prestigecarcompany.co.uk/> Prestige Super Car Sales

...

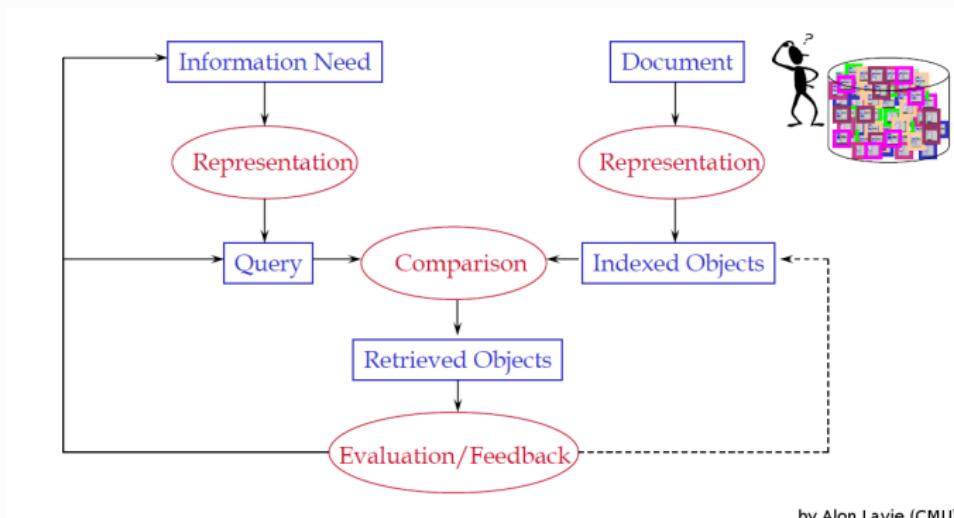
[More videos for black fast jaguar »](#)

Google search (contd)

- What is Google's IR system doing?
 - ◊ Finding pages that contain the words in the query.
- How does it rank these pages?
 - ◊ By “relevance” to the query.
- How does it do it so fast?
 - ◊ By clever indexing (and a lot of hardware!)

Information Retrieval: the task

Text Retrieval: find documents that are “relevant” to a user query.



by Alon Lavie (CMU)

- **Given:** a large, static document collection
- **Given:** an information need (keyword-based query)
- **Task:** find all and only documents relevant to query

Information Retrieval: the task

Typical IR systems:

- Search a set of abstracts
- Search newspaper articles
- Library search
- Search the Web

Typically:

- method more statistics than 'language'
- but the object to retrieve (and process) is language

- How can I formulate a query?
 - ◊ query type: normally keywords, could be natural language
- How are the documents represented?
 - ◊ indexing
- How does the system find the best-matching document?
 - ◊ retrieval model
- How does the system find it *efficiently*?
- How are the results presented to me?
 - ◊ unsorted list, ranked list, clusters
- How do we know whether the system is any good?
 - ◊ evaluation

Reading

- Baeza-Yates and Ribeiro-Neto, Modern Information Retrieval. New York: ACM Press, 1999.
- C. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- I.H. Witten, A. Moffat and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.