Knowledge-based Information Retrieval with Wikipedia

- Koru
- Wikipedia Link-based Measure
- Wikification

The University of Waikato | New Zealand
Limitations of search engines

- “Search is not solved”
- Current search engines
  - don’t understand documents
  - don’t understand queries
Knowledge-based information retrieval

- Consult an external knowledge base
  - to find out what these characters mean
  - and proactively do stuff with them

- A fairly obvious, compelling idea
  - But one that hasn’t worked out

- We haven’t had the right knowledge base
  - Computers aren’t accurate enough
  - Humans aren’t quick enough
Wikipedia | as a knowledge base

- What topics/concepts are there?
  - ~2 million articles and categories
- How are topics referred to?
  - ~5 million titles, redirects and anchors
- How do topics relate to each other?
  - ~60 million article and category links

**Rugby union**

For other uses, see Rugby (disambiguation).

Rugby union (short for rugby union football and often referred to as simply rugby, to a lesser extent football, or union in countries familiar with rugby union and rugby league), is an outdoor sport played by teams of 15 players with an oval ball. It is one of the two main codes of rugby football, the other being rugby league. There is also a quicker seven-a-side variation called rugby sevens, which exists in both forms.

Overview

Main article: Playing rugby union

An adult-level rugby union match lasts for 80 minutes, consisting of two halves of 40 minutes each...
Wikipedia | as a knowledge base

**WordNet**
118,000 synsets

**ResearchCyc**
300,000 concepts
20 years
$\ldots volunteer\ldots$ $\ldots$ volunteer\ldots
1 language

**Wikipedia**
2,000,000 articles
7 years
Almost free
250 languages
2004 U.S. presidential election controversy and irregularities

George W. Bush

Dubya
Shrubya
Thief in chief

Presidents of the United States

Current national leaders

George Walker Bush

Iraq War

2004 U.S. presidential election controversy and irregularities

Al-Qaeda

September 11

Wikipedia as a knowledge base

Formal structure
My research goals

“Wikipedia will provide significantly improved retrieval, as it is”

- We don’t need to make it “tidy”
- It’s not a question of sophisticated NLP or AI
- It’s more about HCI

So let's make a search engine that consults Wikipedia, and find out!
Query Topics

American Airlines Security

Query Results

Done
Fla. Mayor Defends Airport Security

... Mayor Defends Airport Security ... s mayor defended the security at Miami ... of dozens of American Airlines workers in a suspected cocaine ... criticized airport security and said airline workers had easily

American Airlines Fined $250,500

... h1> American Airlines Fined ... 000 civil penalty Thursday against American Airlines because ... their badges in high-security areas at Dallas ... that unless security ... security identification display area

Airline Workers Held in Drug Probe

... - Security lapses at Miami International Airport and ... s managing director of security ... The drugs were put aboard American Airlines flights in Colombia ... agents in Puerto Rico arrested an American Airlines

Airline Workers Held in Drug Probe

... - Security lapses at Miami International Airport and ... s managing director of security ... The drugs were put aboard American Airlines flights in Colombia ... agents in Puerto Rico arrested an American Airlines
Iran condemns air strikes, urges Iraq to cooperate with UN

... N. Security Council to take immediate action to ... The Security Council will not lift economic sanctions imposed on

Muslim Women Claim Discrimination

... discrimination complaint after being fired as airport security ... comply when a supervisor at Argenbright Security Inc

O'Hare Closes Terminal for Security

... Hare Closes Terminal for Security ... authorities could search for a passenger who ran past a security

O'Hare Closes Terminal for Security

... Hare Closes Terminal for Security ... authorities could search for a passenger who ran past a security

Iran and Kuwait sign deal to expand security, drug cooperation

... Iran and Kuwait sign deal to expand security ... Wednesday to expand cooperation in security
Airplane ordered back because of security breach

... Airplane ordered back because of security breach ... a flight without going through security checks ... responsibility for that incident rests squarely with the airline ... Donnolley said security guards notified the control tower.

Calif. Airport Briefly Evacuated

... International Airport on Friday after a man in a business suit ran past a security checkpoint ... which contains Southwest Airlines ... checkpoint to the airline gates for about two hours so security officers could ... The man ignored a security officer.

Airports Lack Security at Bag Claim

... Airports Lack Security at Bag Claim ... perhaps more worrisome for security ... But security at the end of a trip ... said United Airlines spokesman Joe Hopkins.

Report: Jet carrying former U.S. president involved in security

... president involved in security ... President George Bush was involved in a security breach ... was aware of the security breach ... A Qantas spokesman said the airline had taken steps.
security
AND
(air carrier OR airline company OR airline industry OR flight company OR modern aviation OR passenger aircraft)....)
AND
(America OR American OR American continent)....)
Israel increases security at airports following threats...

Israel is adding security personnel at its airports... including airlines... Since the American activity in Afghanistan and Sudan, security... cannot alert security to every metal object.
Israel increases security at airports following threats...

Israel is adding security personnel at its 11 airports, including American Airlines...

Since the American activity in Afghanistan and Sudan, security cannot alert security to every metal object.

Larceny Up at Busiest Airport

Most airport security is focused on major breaches, but Security is through the Federal Aviation Administration. The airlines do report mishandled baggage... a group of baggage handlers at American Airlines in ...

Feds Faulted in Airport Confusion

... diminish her role as an airline safety advocate ... lax airport security and disregard ... America West Airlines and checked her bag at noon ... spokeswoman for America West ... said the airline complied with FAA.

Airport Security Tightened in Miami

... Airport Security Tightened in Miami ... The Miami airport is tightening security after two ... the airport a model of security ... checkpoints staffed by airport security personnel ... many of them American Airline.
Israel increases security at airports following threats

... Israel is adding security personnel at its ... including airlines ... Since the American activity in Afghanistan and Sudan, security ... cannot alert security to every metal object.

Larceny Up at Busiest Airport

... Most airport security is focused on major breaches ... Security is through the Federal Aviation Administration ... The airlines do report mishandled baggage ... a group of baggage handlers at American Airlines in

Feds Faulted in Airport Confusion

... diminish her role as an airline safety advocate ... lax airport security and disregard ... America West Airlines and checked her bag at noon ... spokeswoman for America West ... said the airline complied with FAA.

Airport Security Tightened in Miami

... Airport Security Tightened in Miami ... The Miami airport is tightening security after two ... the airport a model of security ... checkpoints staffed by airport security personnel ... many of them American Airline.

Israel increases security at airports following threats

24 September 1998

Israel increases security at airports following threats

LOD, Israel (AP)

Israel is adding security personnel at its international airport in response to threats by Islamic militants to attack Israeli and U.S. targets, including airlines, officials said Monday.

Threats against U.S. and Israeli targets emerged after last week’s U.S. air strikes in Sudan and Afghanistan that were aimed at Osama bin Laden, a Saudi millionaire and Islamic militant linked to the bombings of two U.S. embassies in East Africa this month.

On Sunday, a leading Muslim activist close to bin Laden said Islamic militants are preparing to retaliate for the air strikes by targeting U.S. and Israeli strategic sites and airliners.

Pini Shis, a spokesman for Israel’s Airport Authority, said more security workers had been added at Ben Gurion International Airport near Tel Aviv, and that security was also being tightened at airfields in Eilat and Tel Aviv.
Israel increases security at airports following threats

... Israel is adding security personnel at its ... including airlines ... Since the American activity in Afghanistan and Sudan, security ... cannot alert security to every metal object

Larceny Up at Busiest Airport

... Most airport security is focused on major breaches ... Security is through the Federal Aviation Administration ... The airlines do report mishandled baggage ... a group of baggage handlers at American Airlines in

Feds Faulted in Airport Confusion

... diminish her role as an airline safety advocate ... lax airport security and disregard ... America West Airlines and checked her bag at noon ... spokeswoman for America West ... said the airline complied with FAA

Airport Security Tightened in Miami

... Airport Security Tightened in Miami ... The Miami airport is tightening security after two ... the airport a model of security ... checkpoints staffed by airport security personnel ... many of them American Airline

15 April 1999

Feds Faulted in Airport Confusion

COLUMBUS, Ohio (AP) -- A former U.S. Transportation Department official who became a top critic of government airline safety policy today blamed airport officials for turmoil that followed her attempt to put an unaccompanied suitcase on a plane.

Mary Schiavo said on NBC's "Today" show that she was sorry for any inconvenience caused when Port Columbus International Airport called the bomb squad and closed a runway for four hours Friday. "But it was done by the airport with the full knowledge of what was going on," she said.

WMCH-TV reported that it had told an airport official Friday morning that it was doing a story on airport security. The station said in a statement Sunday that Ms. Schiavo, working with the station, had checked in a bag for a flight as part of the story, and that airport officials were made aware of the existence of the piece of luggage at around the time the flight was to have taken off.
Wikipedia matches query terminology extremely well.

Recognition and expansion of topics improves retrieval.

Recognition of topics modifies query behavior.

Related topics need further investigation.

Extraction of thesaurus terms is inaccurate.

rugby world cup vs. (rugby world cup OR rwc OR "web ellis cup") rugby world cup vs. "rugby world cup"
What now

We need to improve how topics and the relations between them are extracted

- Semantic Relatedness
- Wikification
Semantic Relatedness

Given any two terms, what is the strength of the semantic relation between them?

- Highly useful
  - AI, data mining, IR, NLP
- But subjective

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Love</td>
<td>Sex</td>
<td>6.77</td>
</tr>
<tr>
<td>Tiger</td>
<td>Cat</td>
<td>7.35</td>
</tr>
<tr>
<td>Tiger</td>
<td>Tiger</td>
<td>10.00</td>
</tr>
<tr>
<td>Book</td>
<td>Paper</td>
<td>7.46</td>
</tr>
<tr>
<td>Computer</td>
<td>Keyboard</td>
<td>7.62</td>
</tr>
<tr>
<td>Computer</td>
<td>Internet</td>
<td>7.58</td>
</tr>
<tr>
<td>Plane</td>
<td>Car</td>
<td>5.77</td>
</tr>
<tr>
<td>Stock</td>
<td>Life</td>
<td>0.92</td>
</tr>
<tr>
<td>Television</td>
<td>Radio</td>
<td>6.77</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Semantic Relatedness | with Wikipedia

Scale and structure

- GBs of text
- millions of articles
- hundreds of thousands of categories

- Two techniques have been developed already

<table>
<thead>
<tr>
<th></th>
<th>WikiRelate!</th>
<th>Explicit Semantic Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19% - 48%</td>
<td>75%</td>
</tr>
</tbody>
</table>
Wikipedia has an extremely rich hyperlink structure that has been ignored so far.

Global warming

Global warming is the increase in the average temperature of the Earth's near-surface air and oceans since the mid-twentieth century, and its projected continuation.

The average global air temperature near the Earth's surface increased 0.74 ± 0.18 °C (1.33 ± 0.32 °F) during the hundred years ending in 2005. The Intergovernmental Panel on Climate Change (IPCC) concluded that most of the observed increase in globally averaged temperatures since the mid-twentieth century is very likely due to the observed increase in anthropogenic (man-made) greenhouse gas concentrations via an enhanced greenhouse effect. Natural phenomena such as solar variation combined with volcanic eruptions probably had a small warming effect from pre-industrial times to 1950 and a small cooling effect from 1961 to 1990.

## Semantic Relatedness | evaluation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>WikiRelate</th>
<th>ESA</th>
<th>WLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miller &amp; Charles</td>
<td>45%</td>
<td>73%</td>
<td>70%</td>
</tr>
<tr>
<td>Rubenstein &amp; Goodenough</td>
<td>52%</td>
<td>82%</td>
<td>64%</td>
</tr>
<tr>
<td>WordSimilarity 353</td>
<td>49%</td>
<td>75%</td>
<td>69%</td>
</tr>
</tbody>
</table>

WikiRelate < WLM < ESA
How do we accurately cross-reference documents with Wikipedia?

Rugby union

From Wikipedia, the free encyclopedia

For other uses, see Rugby (disambiguation).

Rugby union (short for rugby union football and often referred to as simply rugby, to a lesser extent football, or union in countries familiar with rugby union and rugby league), is an outdoor sport played by teams of 15 players with an oval ball. It is one of the two main codes of rugby football, the other being rugby league. There is also a quicker seven-a-side variation called rugby sevens, which exists in both forms.
Wikipedia’s links provide a huge vocabulary of which terms can resolve to which concepts.

“Six central banks, including the Bank of England, have cut interest rates by half a percentage point in an effort to steady the faltering global economy.”
For every link in Wikipedia, a human author has manually chosen the correct destination.

“Six central banks, including the Bank of England, have cut interest rates by half a percentage point in an effort to steady the faltering global economy.”

“The story begins on the banks of the Rio Negro in the Central Amazon. A party of scientists is embarking on a voyage which they hope will provide answers to a five hundred year old mystery.”

recall 96% | precision 98%
Wikification | selecting relevant concepts

- Wikipedians do not link to every single article
  - only ones that readers would want to investigate

  “Six central banks, including the Bank of England, have cut interest rates by half a percentage point in an effort to steady the faltering global economy.”

  “The story begins on the banks of the Rio Negro in the Central Amazon. A party of scientists is embarking on a voyage which they hope will provide answers to a five hundred year old mystery.”

recall 74% | precision 74%
What next?

- Explore applications for Wikification
  - Topic Indexing
  - Document Clustering
  - Document Summarization

- Revisit Koru
  - Apply semantic relatedness and wikification to knowledge base generation, query expansion, and exploratory search

- Write up!
References

Milne, D., Medelyan, O. and Witten, I. H. Mining Domain-Specific Thesauri from Wikipedia: A case study. In *Proceedings of WI 2006*, Hong Kong.


Websites and Demos

www.cs.waikato.ac.nz/~dnk2
www.nzdl.org/koru
wikipedia-miner.sourceforge.net
www.nzdl.org/wikification