Overview of the TREC 2009 Entity track

Krisztian Balog, Arjen P. de Vries, Pavel Serdyukov, Paul Thomas, Thijs Westerveld

The Eighteenth Text Retrieval Conference (TREC 2009)
Gaithersburg, Maryland, USA, November 2009.

Information need

Airlines that currently use Boeing 747 planes

What is the track about?

Airlines that currently use Boeing 747 planes
Airlines that currently use Boeing 747 planes

Boeing 747

The Boeing 747, initially known as the "Jumbo Jet," is one of the world's most recognizable aircraft and has been a symbol of the golden age of air travel for decades. Its design, with two levels of seating, allowed airlines to offer luxury and economy class seating, catering to a wide range of travelers.

The 747 was developed in the early 1960s in response to a need for a new long-range, wide-body airliner that could carry a large number of passengers over long distances. The first 747 made its maiden flight on February 9, 1969, and entered service in 1970.

Key features of the Boeing 747 include:
- Two levels of seating
- Four engines
- Large cargo capacity
- Economical operation

The Boeing 747 has been a staple of the international airline industry and has been used for a variety of applications, from passenger service to cargo transport.

Airlines using the Boeing 747 today include:

- Malaysia Airlines
- Air France
- Lufthansa
- Delta Air Lines
- Cathay Pacific
- Air Canada
- United Airlines
- Emirates

As technology advances, Boeing continues to update the 747 to meet the demands of modern air travel, with newer models offering improved fuel efficiency and increased passenger capacity.

Related lists
- List of Boeing 747 operators
- Boeing 747 technical specifications
Envisaged interface

Japan Airlines
Japan Airlines International Co., Ltd. is the flag carrier of Japan... The group has a fleet of 279 aircraft, consists of only Boeing aircraft for its long-haul operations, and currently has a fleet of Airbus, Boeing, and McDonnell Douglas aircraft for its short-haul operations.

British Airways
British Airways is the flag carrier airline of the United Kingdom. It is headquartered in Waterside near its main hub at London Heathrow Airport and is the largest airline in the UK based on fleet size, international flights and international destinations.

KLM Royal Dutch Airlines
KLM is the national airline of the Netherlands and is part of Air France-KLM. KLM operates worldwide scheduled passenger and cargo services to more than 90 destinations.

What is an entity?

• Entity = something with a homepage

Related entity finding

• Given
  • Input entity, defined by its name and homepage
  • Type of the target entity (person, organization, or product)
  • Narrative (describing the nature of the relation in free text)
  • Return homepages of related entities

Example

<query>
  <num>7</num>
  <entity_name>Boeing 747</entity_name>
  <entity_URL>clueweb09-en0005-75-02292</entity_URL>
  <target_entity>organization</target_entity>
  <narrative>Airlines that currently use Boeing 747 planes.</narrative>
</query>

An answer record

• HP Homepages of the entity (up to 3)
• WP Wikipedia page of the entity
• NAME String answer that represents the entity concisely
• SUPPORT Supporting documents (up to 10)
Example

Related tasks

- QA list questions
- They don't have a relationship component
- No homepages for entities
- Entity ranking @INEX
- Different task
- Only entities in Wikipedia
- No restrictions on entity types

Informal reformulation of the task for “TREC regulars”

- Mixture of QA + Homepage finding
- Return homepages of answers to questions

Outline

- Introduction
- Topic development
- Judgments
- Results
- Approaches
- Outlook

Data collection

- Category B subset of ClueWeb

Topic development

"The hardest track I have ever developed topics for."

(Ian Soboroff)
Topic development

- Main problem: No proper tools exist today
- Related problems
  - Not many entities have homepage in CatB
  - What is a homepage, anyway?
  - Trivial topics
  - QA list topics could not be reused

Hunt for entities on the Web, then map them back to ClueWeb
- Burn rate: 1/5 - 1/10

Assessment procedure

- Decide if HP{1,2,3} are primary or relevant
- Decide if WP is primary or relevant
- Decide if NAME is correct or inexact for something in the record
- Clustering pages and names that belong to the same entity

Metrics

- \textbf{NDCG}$_R$ Normalized discounted cumulative gain at rank R
- \textbf{R} is the number of primaries and relevent for the topic
- \textbf{P@10} Fraction of primary pages in the top 10 results
- \textbf{#rel} Number of relevant pages
- \textbf{#pri} Number of primary pages

Runs

- Max. 4 runs/team allowed
- 13 participating teams
- 41 runs
  - 33 automatic / 8 manual

Topics

<table>
<thead>
<tr>
<th>Number of primary HPs per topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Target type

- Organization
- Person
- Product

<table>
<thead>
<tr>
<th>Target type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>3</td>
</tr>
<tr>
<td>Person</td>
<td>6</td>
</tr>
<tr>
<td>Product</td>
<td>11</td>
</tr>
</tbody>
</table>

Number of primary HPs per topic
### Results

<table>
<thead>
<tr>
<th>Group</th>
<th>HP (1,2,3)</th>
<th>WP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purdue</strong></td>
<td>126</td>
<td>90</td>
</tr>
<tr>
<td><strong>uogTr</strong></td>
<td>347</td>
<td>73</td>
</tr>
<tr>
<td><strong>CAS</strong></td>
<td>80</td>
<td>63</td>
</tr>
<tr>
<td><strong>NICT</strong></td>
<td>99</td>
<td>67</td>
</tr>
<tr>
<td><strong>UAms (Amst.)</strong></td>
<td>198</td>
<td>20</td>
</tr>
<tr>
<td><strong>TUDelft</strong></td>
<td>108</td>
<td>32</td>
</tr>
<tr>
<td><strong>BUPTPRIS</strong></td>
<td>48</td>
<td>14</td>
</tr>
<tr>
<td><strong>UALR_CB</strong></td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td><strong>UIUC</strong></td>
<td>64</td>
<td>13</td>
</tr>
<tr>
<td><strong>Waterloo</strong></td>
<td>55</td>
<td>9</td>
</tr>
<tr>
<td><strong>EceUdel</strong></td>
<td>102</td>
<td>45</td>
</tr>
<tr>
<td><strong>BIT</strong></td>
<td>81</td>
<td>47</td>
</tr>
<tr>
<td><strong>UAms (ISLA)</strong></td>
<td>30</td>
<td>9</td>
</tr>
</tbody>
</table>

• Bootstrapping Wikipedia
• Name variant detection
• Exploiting lists and tables
• Using Google
Some dos and don’ts

- What you should not do
  - Take qrels and report results in your SIGIR paper
  - Topic set is too small
- What you should do
  - Look at results topic by topic
  - What worked and what didn’t

Envisaged interface

Entity track 2010

Pilot task 1: summary generation

Airlines that currently use Boeing 747-400 planes

<table>
<thead>
<tr>
<th>Airlines</th>
<th>Country</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan Airlines International Co., Ltd.</td>
<td>Japan</td>
<td>The group has a fleet of 279 aircraft, consists of only Boeing aircraft for its long-haul operations, and Airbus, Boeing and McDonnell Douglas aircraft for its short-haul operations.</td>
</tr>
<tr>
<td>British Airways</td>
<td>United Kingdom</td>
<td>British Airways is the flag carrier airline of the United Kingdom. It is headquartered in Waterside near its main hub at London Heathrow Airport and is the largest airline in the UK based on fleet size, international flights and international destinations.</td>
</tr>
<tr>
<td>KLM Royal Dutch Airlines</td>
<td>Netherlands</td>
<td>KLM is the national airline of the Netherlands and part of Air France-KLM. KLM operates worldwide scheduled passenger and cargo services to more than 90 destinations.</td>
</tr>
</tbody>
</table>

Entity track 2010

Pilot task 2: aspect identification

Airlines that currently use Boeing 747-400 planes

<table>
<thead>
<tr>
<th>Airlines</th>
<th>Country</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan Airlines International Co., Ltd.</td>
<td>Japan</td>
<td>The group has a fleet of 279 aircraft, consists of only Boeing aircraft for its long-haul operations, and Airbus, Boeing and McDonnell Douglas aircraft for its short-haul operations.</td>
</tr>
<tr>
<td>British Airways</td>
<td>United Kingdom</td>
<td>British Airways is the flag carrier airline of the United Kingdom. It is headquartered in Waterside near its main hub at London Heathrow Airport and is the largest airline in the UK based on fleet size, international flights and international destinations.</td>
</tr>
<tr>
<td>KLM Royal Dutch Airlines</td>
<td>Netherlands</td>
<td>KLM is the national airline of the Netherlands and part of Air France-KLM. KLM operates worldwide scheduled passenger and cargo services to more than 90 destinations.</td>
</tr>
</tbody>
</table>
Questions?

http://ilps.science.uva.nl/trec-entity/guidelines/plans-for-2010/