living knowledge

http://livingknowledge-project.eu/

SEVENTH FRAMEWORK PROGRAMME

Search Computing and Social Media Workshop

NEM Summit 2011, Torino, September 28, 2011

MAKING CONTENT DIVERSITY AN ASSET IN SEARCH

Claudia Niederée
L3S Research Center, Hannover
Research Manager of LivingKnowledge Project

MOTIVATION

Bias in the use of images

article | discussion | view source | history

Global warming

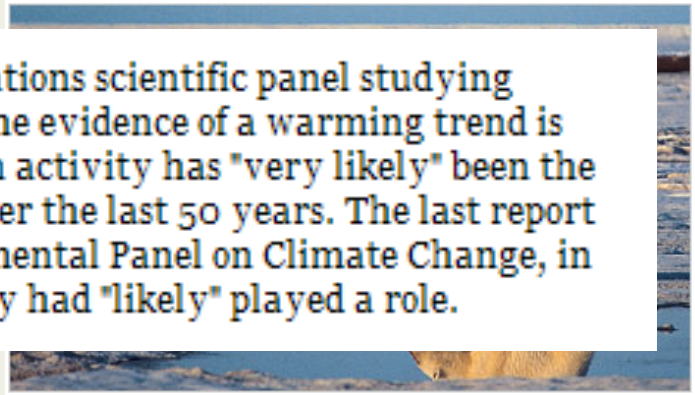
From Wikipedia, the free encyclopedia

This article is about the current period of increasing global tempera see Palaeoclimatology and Geologic temperature record.

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 increa 0.74 ± 0.18 °C (1.33 ± 0.32 °F) during the hundred years ending in 2005.[1] The Intergovernmental Panel on Climate Change (IPCC) concludes "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"[1] via the greenhouse effect. Natural phenomena such as solar variation combined with volcanoes probably had a small warming effect from pre-industrial times to 1950 and a small cooling effect from 1950 onward.[2][3]

„global warming“



Subhankar Banerjee/Associated Press

On Feb. 2, 2007, the United Nations scientific panel studying climate change declared that the evidence of a warming trend is "unequivocal," and that human activity has "very likely" been the driving force in that change over the last 50 years. The last report by the group, the Intergovernmental Panel on Climate Change, in 2001, had found

AN INCONVENIENT TRUTH

bias? interes behind inform complete opi overview?

Posted on **Monday, June 26, 200**

Before we get too hyped up a propaganda.

Here is information from an ar warming:

http://www.reason.com/rb/rb0

living knowledge

http://livingknowledge-project.eu/

Diversity and bias in the Web today

Web today

- diverse content provided by multitude of stakeholders
- freedom of publication + democratization of publication process
- further strengthened by Social Web
  - high diversity in available content
  - high volumes of user generated content
  - high user involvement
  - more opinionated content

see e.g. Study by Universal McCann from March 2008*

- 184 million WW have started a blog | 26.4 US
- 346 million WW read blogs | 60.3 US
- 77% of active Internet users read blogs

- ... but:
  - discovery of diverse positions on a topic by chance
  - no systematic support to explore the diversity
  - risk of biasing

*http://www.universalmccann.com/Assets/UM%20Wave%203%20final_20080808141650.pdf

SEVENTH FRAMEWORK PROGRAMME

living knowledge

http://livingknowledge-project.eu/

THE LIVINGKNOWLEDGE PROJECT

- FET project in the area „ICT forever yours“
- Start: February 2009, 36 months
- Interdisciplinary research team
  including Uni Trento, Yahoo! Research,
  MPI, L3S Research Center,
  Uni Southampton, ISI Bangalore, …

- Project goal:
  improve navigation and search in large cross-media datasets and the Web by considering diversity, time, opinions and bias

SEVENTH FRAMEWORK PROGRAMME

Making diversity an asset: Example Technologies



Professional & Private users

Diversity-aware Technology

tangible diversity

Search result diversification (text)

Search result diversification (images)

diversity dimension time

Searching the past present & Future

discourse analysis

opinions

Opinion mining

sentiment analysis for images

use of images

detection of photo montage

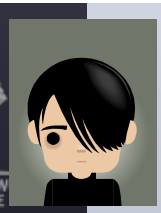
image forensics detection of + tampering

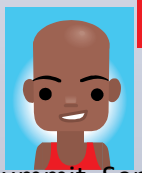
image manipulation impact analysis

knowledge in context

Fact extraction & Fact evolution analysis

Diversity of Content Creators






Opinions

Biased Content

Diverse Content

SEARCH RESULT DIVERSIFICATION*

- general goal: get top k results that are relevant, but not too similar (good coverage)
- balancing between:
  - relevance of query result
  - overlap between query results
- different objective function can be used
- optimization problem, which is NP hard

→ use of approximation algorithm

- Adequate evaluation is a challenge as well

*Ralf Krestel, Peter Fankhauser: **Ranking Web search Results for Diversity,** Journal of Information Retrieval, Springer (under revision)

living knowledge

http://livingknowledge-project.eu/

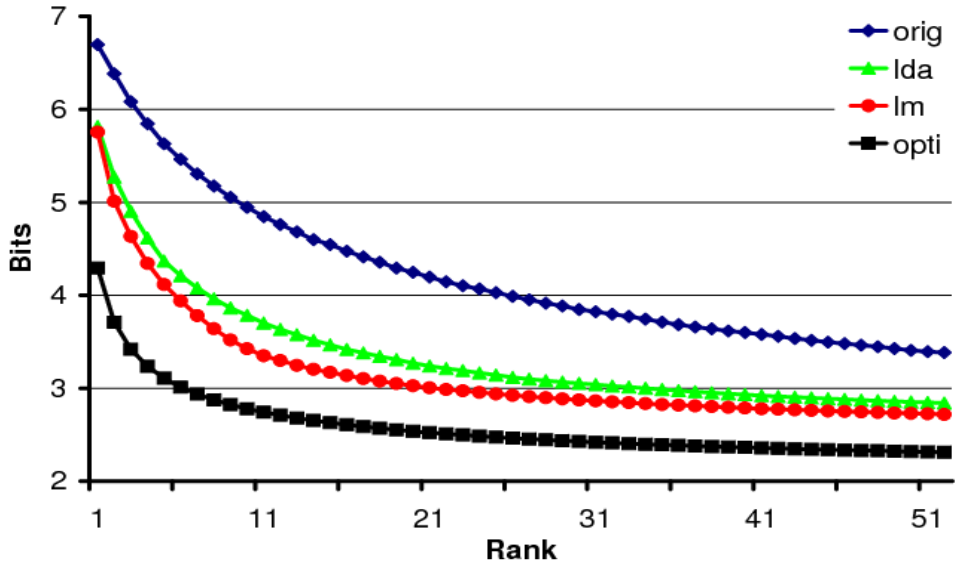
SEVENTH FRAMEWORK PROGRAMME

BIASITY APPROACH

- optimization goal: maximizing relevance and minimizing variance of result set
- use of a greedy algorithm
- experiments with document representations based on a) language model b) topic model (LDA)
- relies on ranking of Web search engine (google, Yahoo!) ; diversification by re-ranking
- use of discount function to punish re-ranking

BIASITY –EVALUATION

- Groundtruth construction based on ambigous Wikipedia titles
- Build language models from page content
- Compare search engine results and diversified results with the ground truth for each query
- Use Kullback-Leibler Divergence to measure coverage
- Use Spearman's rank correlation to measure re-ranking



living knowledge

http://livingknowledge-project.eu/

FUTURE PREDICTOR -> TIME EXPLORER*

- Idea: search in the past, present … and in the future
- uses 20 years coverage of NYT corpus
- builds upon extraction technology, especially extraction of entities and of time-related expressions
- indices for publication dates and content dates

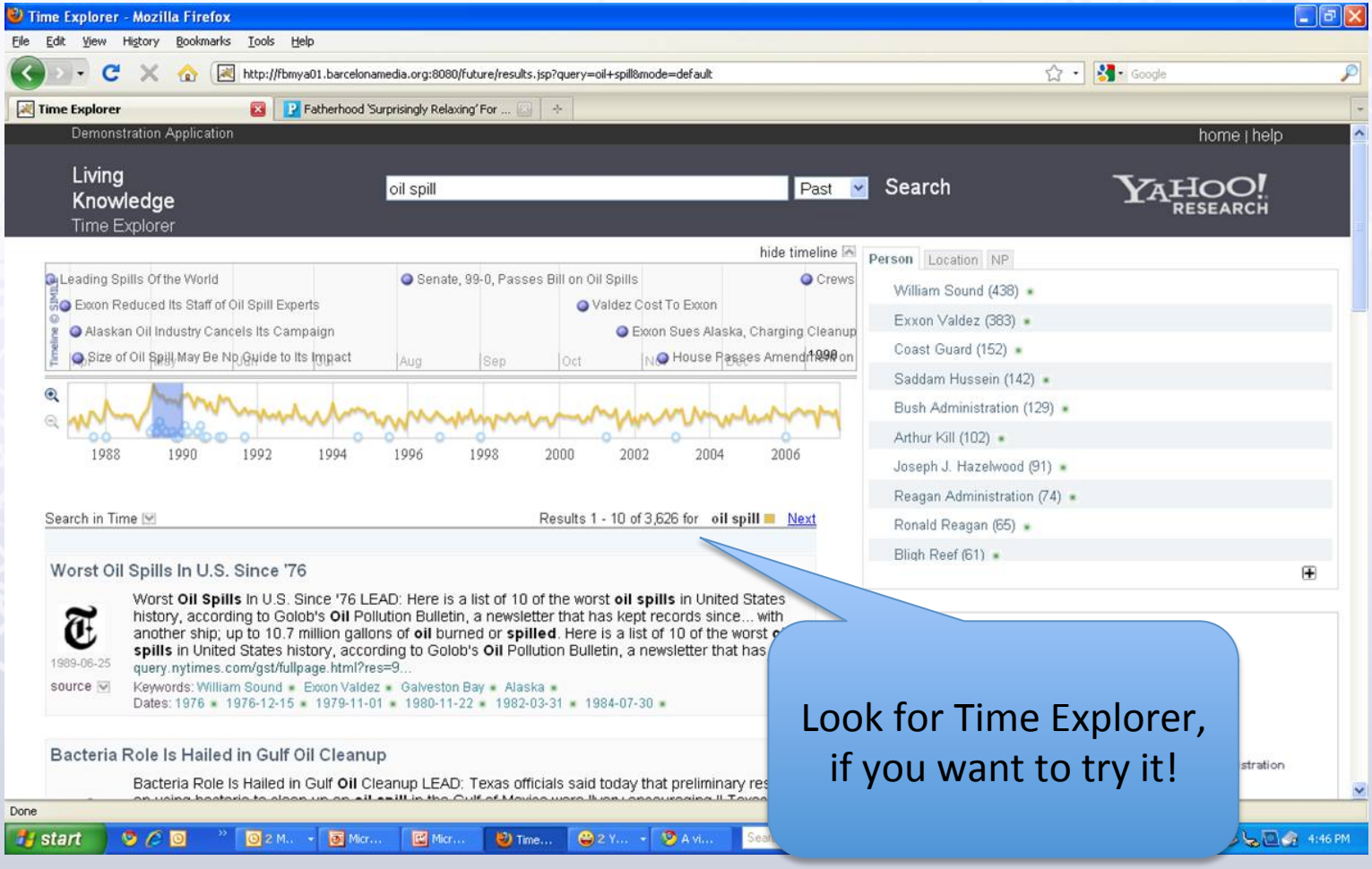
→ enables a wide variety of queries to be answered:

  - understand how a topic/story evolves
  - see the main entities (e.g. persons) related to a topic at different points in time (and their relationship)
  - number of documents on a topic
  - search for statements about the future
  - …

***Michael Matthews et al.: Searching through time in the New York Times, HCIR Challenge 2010**

SEVENTH FRAMEWORK PROGRAMME

TIME EXPLORER – USER INTERACTION



living know ledge

http://livingknowledge-project.eu/

SEVENTH FRAMEWORK PROGRAMME

THANKS!

Contact: Claudia Niederée, niederee@l3s.de

- a lot has been achived, ..
- … but there is also still a lot to do:
  - e.g. automatic detection of bias in text and the use of images
  - e.g. adequate visualization of diverse content

More Info?

- **Project Page:** http://livingknowledge-project.eu/
- Many **publications** on individual results
- **Diversity Engine:** Testbed containing
  - about 50 components for annotating document collections with a wide range natural language processing and image analysis tools
  - provides methods for indexing, searching, and visualizing these an-notations using the Solr search engine
  - Check out at http://sourceforge.net/p/diversityengine