

living  
know  
ledge

<http://livingknowledge-project.eu/>



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

## Global warming

From Wikipedia, the free encyclopedia

*This article is about the current period of increasing global temperature. For the study of past climate, 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 increased by  $0.74 \pm 0.18 \text{ }^\circ\text{C}$  ( $1.33 \pm 0.32 \text{ }^\circ\text{F}$ ) during the hundred years ending in 2005.<sup>[1]</sup> 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"<sup>[1]</sup> 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.<sup>[2][3]</sup>

Bias in the use of images

„global warming“



bias? interest behind information? complete opinion overview?



## AN INCONVENIENT TRUTH

Person

Posted on Monday, June 26, 2006

Before we get too hyped up about global warming, let's look at the propaganda.

Here is information from an article about global warming:

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

NET Summit, September 2011

melting of the Antarctic and

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 that humanity had "likely" played a role.

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 that humanity had "likely" played a role.



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

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

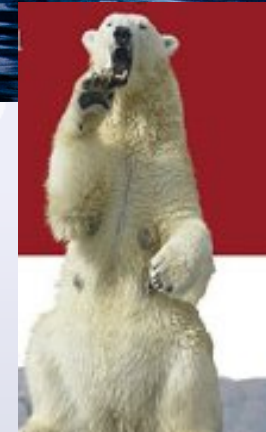
- 184 million WW have started a blog | 26.4 US
- 346 million WW read blogs | 60.3 US
- 77% of active Internet users read blogs

\*[http://www.universalmccann.com/Assets/UM%20Wave%203%20final\\_20080808141650.pdf](http://www.universalmccann.com/Assets/UM%20Wave%203%20final_20080808141650.pdf)



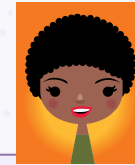
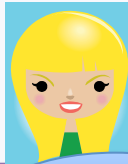
# THE LIVINGKNOWLEDGE PROJECT

- 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



# Making diversity an asset: Example Technologies

Professional  
& Private users



Diversity-aware  
Technology

Search result  
diversification  
(text)

tangible  
diversity

Search result  
diversification  
(images)

diversity  
dimension  
time

Searching the past  
present & Future

discourse  
analysis

opinions

sentiment  
analysis for  
images

use of  
images

detection of  
photo montage

Opinion  
mining

image forensics  
detection of +  
tampering

image  
manipulation  
impact analysis

knowledge  
in context

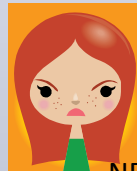
Fact extraction &  
Fact evolution  
analysis

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)

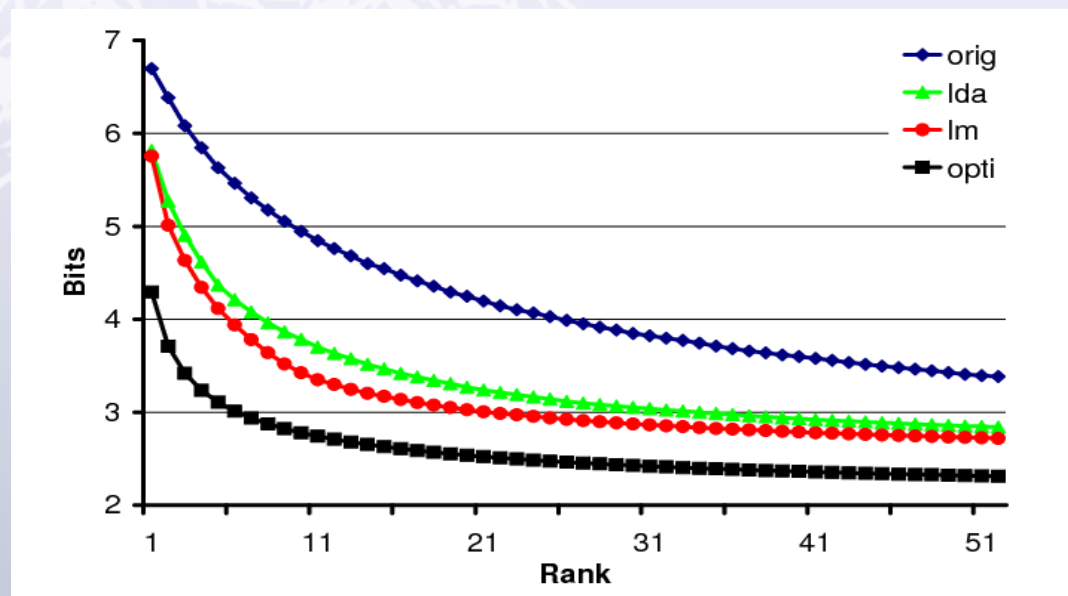
## BIASITY APPROACH

- 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 ambiguous 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





## 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

# TIME EXPLORER – USER INTERACTION

Time Explorer - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://fbmya01.barcelonamedia.org:8080/future/results.jsp?query=oil+spill&mode=default

Time Explorer

Demonstration Application home | help

Living Knowledge Time Explorer

oil spill Past Search

YAHOO! RESEARCH

hide timeline

Leading Spills Of the World

Senate, 99-0, Passes Bill on Oil Spills

Exon Reduced Its Staff of Oil Spill Experts

Valdez Cost To Exon

Alaskan Oil Industry Cancels Its Campaign

Exxon Sues Alaska, Charging Cleanup

Size of Oil Spill May Be No Guide to Its Impact

House Passes Amendment on

Person Location NP

- William Sound (438)
- Exxon Valdez (363)
- Coast Guard (152)
- Saddam Hussein (142)
- Bush Administration (129)
- Arthur Kill (102)
- Joseph J. Hazelwood (91)
- Reagan Administration (74)
- Ronald Reagan (65)
- Bligh Reef (61)

Search in Time Results 1 - 10 of 3,626 for oil spill Next

Worst Oil Spills In U.S. Since '76

Worst **Oil Spills** In U.S. Since '76 LEAD: Here is a list of 10 of the worst **oil spills** in United States history, according to Golob's **Oil Pollution Bulletin**, a newsletter that has kept records since... with another ship; up to 10.7 million gallons of **oil** burned or **spilled**. Here is a list of 10 of the worst **spills** in United States history, according to Golob's **Oil Pollution Bulletin**, a newsletter that has...

1989-06-25 source

Keywords: William Sound • Exxon Valdez • Galveston Bay • Alaska •

Dates: 1976 • 1976-12-15 • 1979-11-01 • 1980-11-22 • 1982-03-31 • 1984-07-30 •

Bacteria Role Is Hailed in Gulf Oil Cleanup

Bacteria Role Is Hailed in Gulf **Oil** Cleanup LEAD: Texas officials said today that preliminary res...

start 2 M... Micr... Micr... Time... 2 Y... A vi... 4:46 PM

Look for Time Explorer, if you want to try!

# THANKS!

**Contact:** Claudia Niederée,  
[niederee@l3s.de](mailto:niederee@l3s.de)

- a lot has been achieved, ..
- ... 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 annotations using the Solr search engine
  - Check out at <http://sourceforge.net/p/diversityengine>