CURRENT APPROACHES TO SEARCH RESULT DIVERSIFICATION

Motivation

The Web grows, the number of relevant results grows as well

Search engine users look only at top few documents

They should be a *good* sample of the entire relevant set

Outline

Diversification of Web results: problem definition

A Framework for search result diversification

- Relevance Functions
- Similarity measures
- Objective functions

Datasets and evalution techniques

Room for improvement

Diversifying Web Search Results

Levels of diversity in Web Search

- Ambiguous queries: different senses
- Clear queries: different aspects/subtopics

Problem: find the subset with the *k* most **relevant** and **diverse** results

In a ranked list:

- Top-k docs results are relevant
- i-th result should be novel compared to the i-1 previous docs

Diversifying Web Search Results

Different types of diversity exists

- Topic, Opinion, Genre, Document type, Time, Conflicting info, ...

Different applications can benefit from *result diversification*

- Web Search
- News
- Blogs
- Product Search
- ...

Trade-off relevance/novelty

Finding the optimal set of items which is both relevant and diverse

- Relevance measure
- Similarity (diversity) measure
- Diversification objective (trade-off)
    - NP-hard problem
    - Use greedy algorithms
    - Compute an aproximation

Relevance measure

All systems work on top-k items ordered by a relevance measure

For both full text and structured datasets

Different measures can be used to identify such set:

- Language models [1,8]
- Vector space [2] BM25 [8]
- KL-divergence [4]
- ...

Similarity measures

- Semantic Distance (Textual similarity)
    - Cosine sim
    - Jaccard sim [1]
    - Euclidean distance [2]
- Categorical distance
    - Tree distance based on taxonomies [1] [3]
    - Order of attributes to be diversified [5]
- Novel Information
    - KL Divergence [2]
- Query reformulations from WSE + collection statistics [8]

- No measure exploits genre, sentiment, or other diversity types

Objective functions

Combining relevance and diversity

Find the optimal set of items which is relevant and diverse

Proposed objective functions:

- Max-sum [1]: weighted sum
- Max-min [1]: min relevance and dissimilarity
- Average dissimilarity [1]: adds to the relevance the avg dissimilarity
- Max-sum of max-score [5]: max diversity after max relevance
- Max-product [4]: select i-th results by relevance*dissimilarity(i,1..i-1)
- Categorical diversification [3]: covered categories
- Probability mixture model [8]: weighted sum

The problem is NP-hard

Aproximations use **on-line** greedy algorithms

Datasets

Main distiction is between full text vs structured datasets

Full text:

- Top k docs from commercial search engines [3]
- TREC Interactive [4]
- TREC Web – diversity task [8]

Structured data:

- Yahoo! Autos [5]
- DB [2]
- IMDB [7]
- Syntetic datasets [2]

Ground truth:

- Wikipedia disambiguation pages [1]
- Amazon Mturk [3]

Evaluation Measures

New diversity aware measures are defined for IR tasks only

- alpha-NDCG [6]: relevance based on subtopics covered in the query and contained in previous results
- S-Precision, S-Recall (aka novelty [1]), WS-Precision [4]
- NDCG-IA MAP-IA MRR-IA [3]: user intent

DB search

- goodness of the aproximation compared to the optimal result
- Efficiency
- alpha-NDCG-W [7]: judges the ranking of **query interpretations**
- WS-Recall [7]: different importance of sub-topics

TREC 2009 Web Track

Diversity Task

- Return a ranked list that provides complete coverage for a query
- Avoiding redundancy in the result list

Subtopics, each related to a different user need

- For each subtopic, assessors make a binary relevance judgment

Measures:

- α-nDCG
- MAP-IA
- give no credit to duplicate and near-duplicate documents

http://plg.uwaterloo.ca/~trecweb/

Example topic

Topic: physical therapist

Subtopics (not given!):

- What does a physical therapist do?
- Where can I find a physical therapist?
- Therapy cost per hour
- Required Training
- American Physical Therapy Association
- Salary
- Difference between a occupational therapist and a physical therapist
- Required education

Topical diversity

Possible next steps

Algorithms

- Off-line steps to simplify the on-line optimization step
    - Relevance functions focusing on diversity (no re-ranking)
    - [5] proofs that inverted indexes can not do that
- Other diversity notions: similarity measures not based on content
    - Opinion, topic, genre, time, ...
    - Combine different notions in one measure

Interaction

- What diversity the user expects?

Benchmarks

- TREC is producing a topical-diversity benchmark
- One corpus for each notion of diversity should be created

References

[1] Gollapudi, S., Sharma, A.: An Axiomatic Approach for Result Diversification. In: WWW '09

[2] Jain, A., Sarda, P., Haritsa, J.R.: Providing Diversity in K-Nearest Neighbor Query Results. In: PAKDD04

[3] Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying Search Results. In: WSDM09

[4] Zhai, C.X., Cohen, W.W., Laerty, J.: Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In: SIGIR03

[5] Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., Yahia, S.A.: Efficient Computation of Diverse Query Results. In: ICDE08

[6] Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Buettcher, S., MacKinnon, I.: Novelty and Diversity in Information Retrieval Evaluation. In: SIGIR08

[7] Demidova, E., Fankhauser, P., Zhou, X., Nejdl, W.: DivQ: Diversification for Keyword Search Over Strucutred Databased. In: SIGIR10

[8] Santos, R., Macdonald, C., Ounis, I.: Exploiting Query Reformulation for Web Search Result Diversification. In: WWW10

living
know
ledge

http://livingknowledge-project.eu/

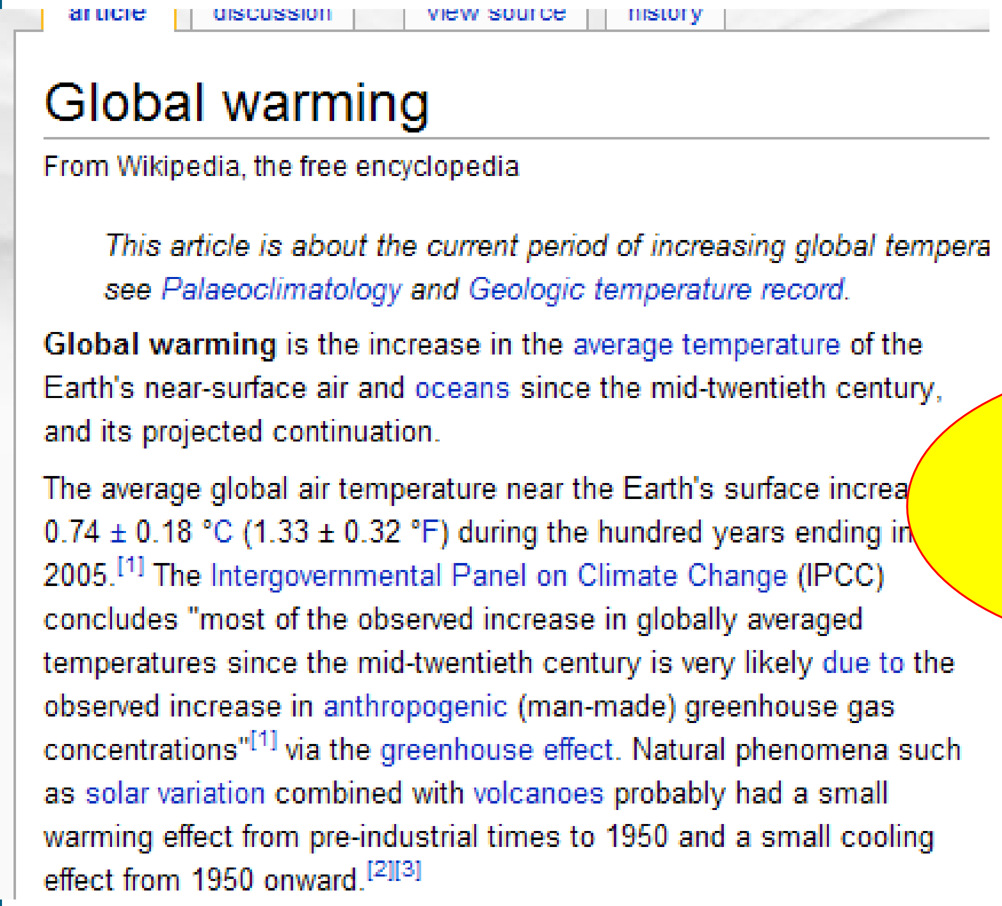
SEVENTH FRAMEWORK PROGRAMME

Living Knowledge Project

PROJECT OVERVIEW

Gianluca Demartini

Motivation



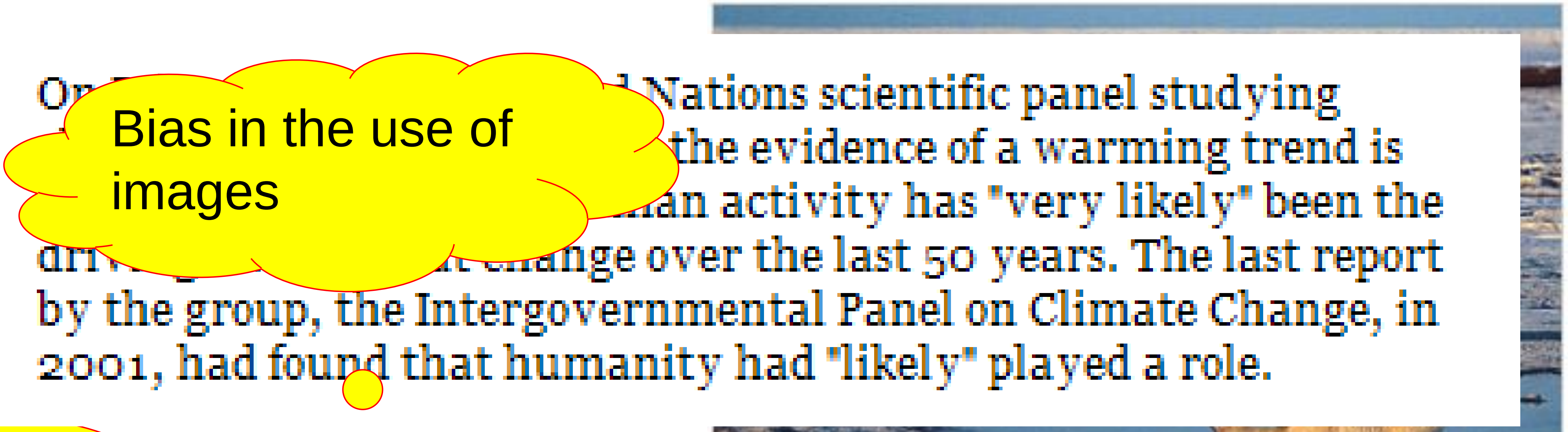
Global warming

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.

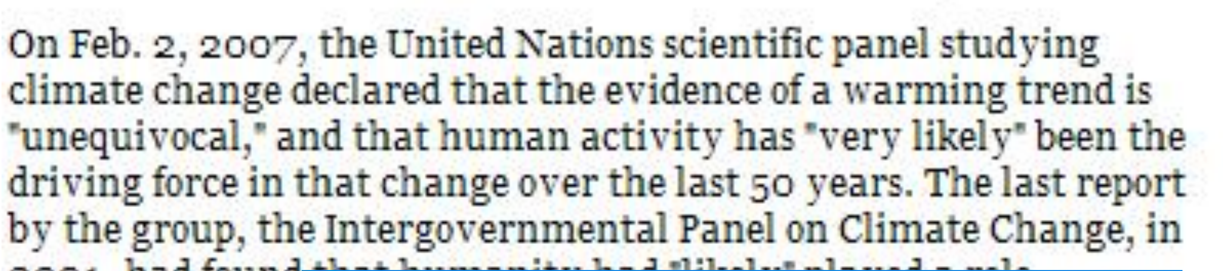
The average global air temperature near the Earth's surface increa 0.74 ± 0.18 °C (1.33 ± 0.32 °F) during the hundred years ending in 2005.[1] The Intergovernmental Panel on Climate Change (IPCC) concludes "most of the observed increase in globally averaged temperatures since the mid-twentieth century 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]



Bias in the use of images

On F Nations scientific panel studying the evidence of a warming trend is an activity has "very likely" been the dri 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

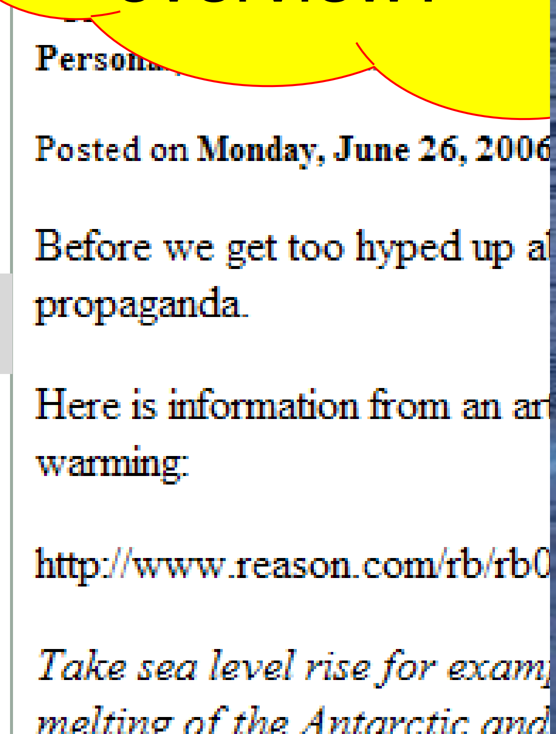


On Feb. 2, 2007, the United Nations scientific panel studying climate change declared that the evidence of a warming trend is "unequivocal," and that human activity has "very likely" been the driving force in that change over the last 50 years. The last report by the group, the Intergovernmental Panel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

„global warming“

AN INCONVENIENT TRUTH

bias? interest behind inform complete opi overview?



Person

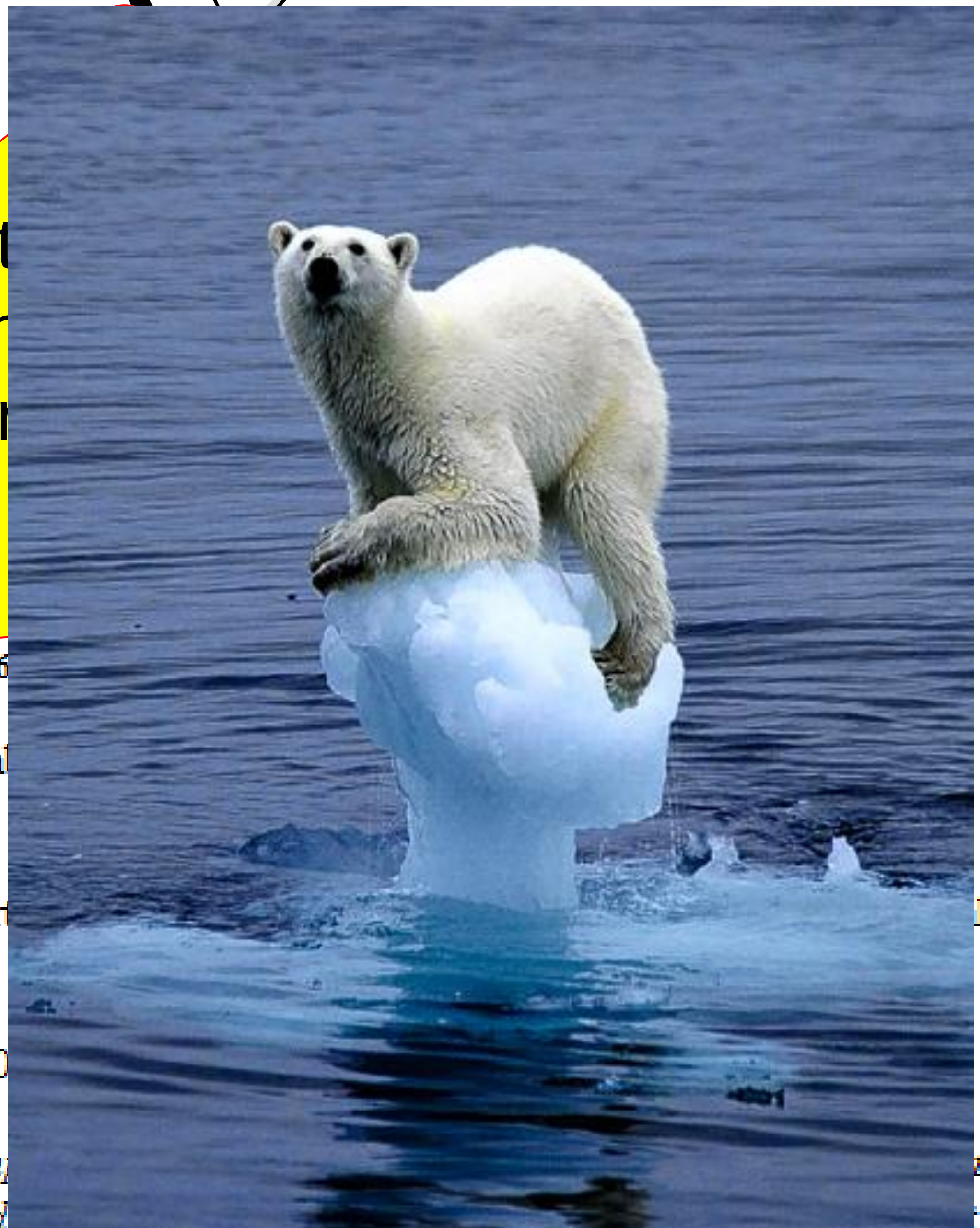
Posted on **Monday, June 26, 2006**

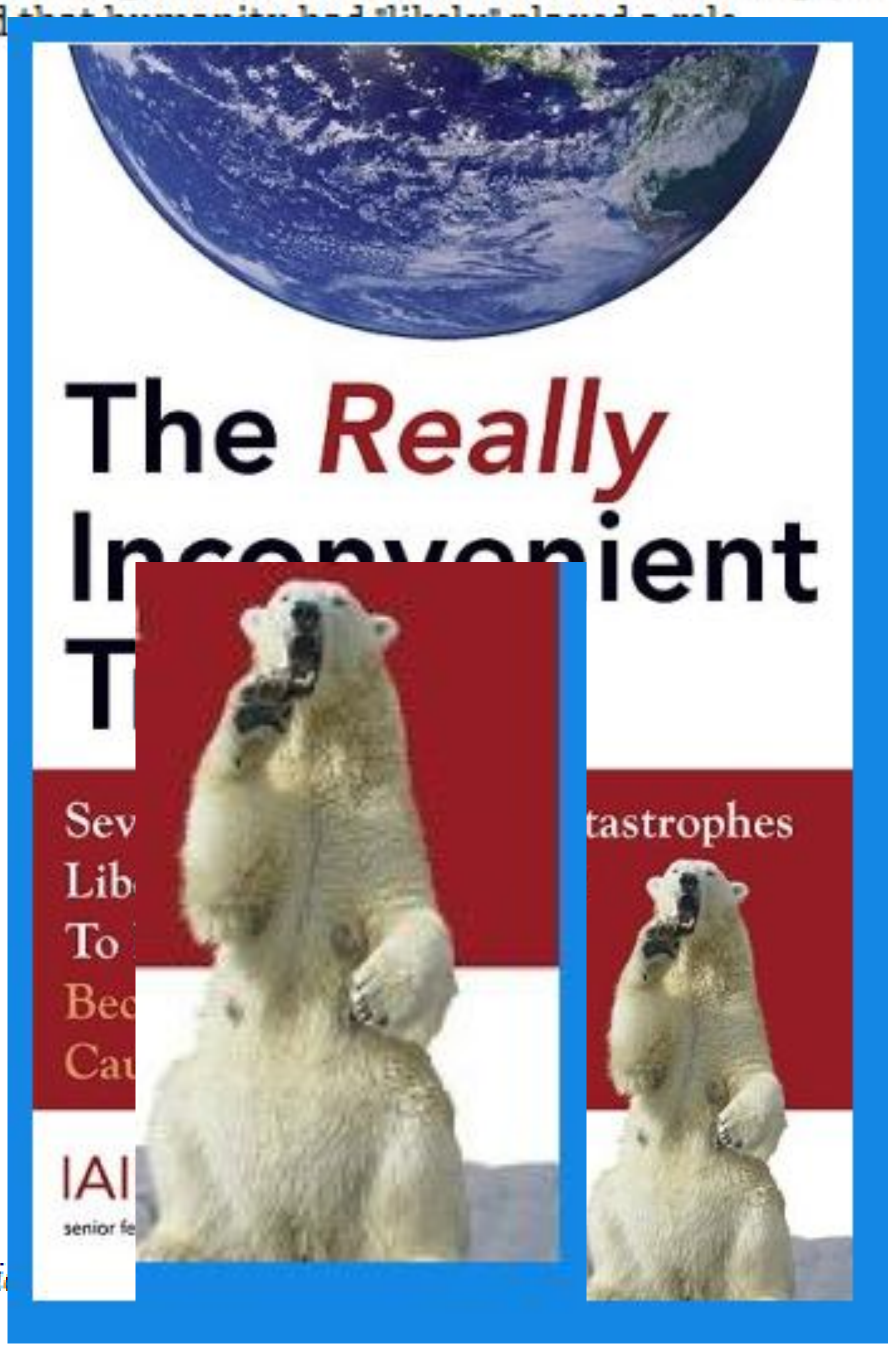
Before we get too hyped up a propaganda.

Here is information from an ar warming:

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

Take sea level rise for examp melting of the Antarctic and





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 Web 2.0
  - 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

LK PROJECT VISION AND OBJECTIVES

Project Vision: make diversity, bias and evolution traceable, understandable and exploitable

Objectives:

- Creating a deep understanding of diversity and how it reflects in content
- Exploring the temporal dimension of knowledge
- Developing methods for detecting bias
- Making bias, diversity and evolution tangible and digestible by a new generation of search technology
- Forwarding the research area and raising awareness and building a community around the RTD

living
know
ledge

http://livingknowledge-project.eu/

SEVENTH FRAMEWORK
PROGRAMME

THANKS