



**Thomas Mandl**  
Information Science  
Universität Hildesheim  
Germany  
mandl@uni-hildesheim.de



## Implementation and Evaluation of a Quality-Based Search Engine



Hypertext 2006  
Odense



## Overview

- Context
- Quality in the Internet
- Link Analysis
- Alternative Methods for Quality Assessment
- AQUAINT Project (Automatic Quality Assessment for Internet Resources)
  - AQUAINT Model
  - Implementation
  - Evaluation

## Lack of Quality on the Internet

- “a large fraction of **low quality** web pages that users are unlikely to read” (Page et al. 1998:2)
- “**False** information abounds, either accidentally or with evil intent” (Weinstein & Neumann 2000)
- “information **quality varies** widely on the Internet” (Zhu & Gauch 2000:288)

## Automatic Quality Assessment is Reality


- Automatic Grading of Essays for College Entry Exams in the USA (Miltasakaki & Kukich 2004)
- Recommendation Systems: human judgements are aggregated and weighted by complex algorithms (Avesani et al. 2005)

## Framework for Definitions of Quality

- Transcendent:** objective and absolute quality, which is universally valid.
- User-oriented:** subjectivity, quality depends on context and situation of the user

*cf. Marchand 1990*

## Link-Analysis



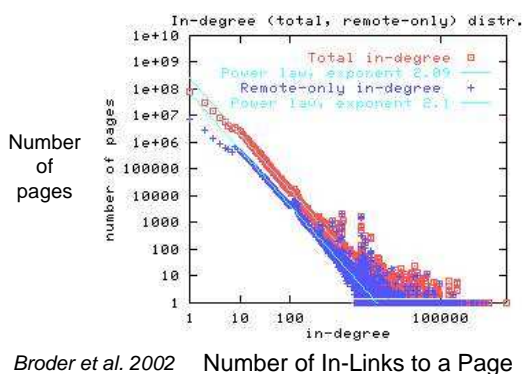
## Link-Analysis: Basic Idea

- Current standard approach to automatic quality assessment
- Basic idea stems from Biblio- or Scientometrics
- Many links to an object support its authority
- Most well known algorithm: PageRank (maybe applied by Google)

## Link-Analysis: PageRank

- The more links pointing to a page, the higher is its authority
- The higher the authority of a page, the more it contributes to the authority of the target page
- Iterative algorithm

## Link-Distribution



## Growth Model

$$\Pi(l(i)) = \alpha \frac{lc(i)}{L} + (1 - \alpha) \frac{1}{U}$$

0.9      0.1

$\Pi(l(i))$     Probability, that new link refers to unit  $i$   
 $lc(i)$        number of in-links of unit  $i$  (Link - Count)  
 $L$             current number of links in the network  
 $U$             current number of units in the network  
 $\alpha$            parameter

α = 0.9 !  
Matthew-Effect!      (PENNOCK ET AL. 2002:3)

## Matthew-Effect

- Jesus said:
- **“For everyone who has will be given more, and he will have an abundance. Whoever does not have, even what he has will be taken from him.”**  
 (Matthew 25:29)

## TREC: Approach

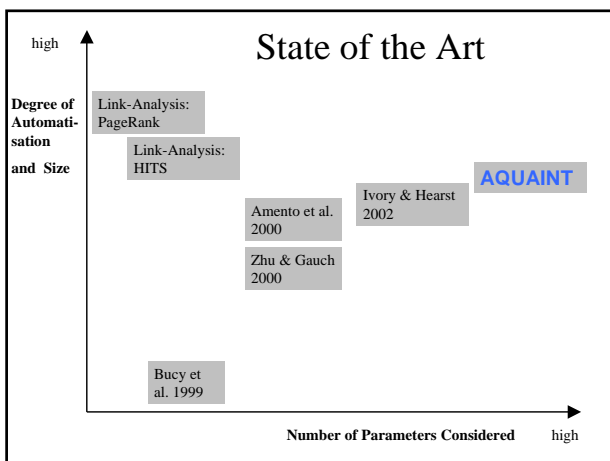
- Text Retrieval Conference
- Test Basis
  - Objects (Documents, ....)
  - Information Requests (Topics)
  - Standard Relevance Assessment
- Starting in 2000: Web Track
  - Different Corpora („web snapshots“)
  - Evaluation of Web Retrieval Algorithms

## Web-Track: Results

- Several groups tested PageRank in the TREC web track
- **Improvement could only be noted for the homepage finding task**

## Link-Analysis

- Link Analysis is insufficient as the only basis for quality assessment
- experimental systems are searching for alternative approaches
- -> **AQUAINT**



## AQUAINT

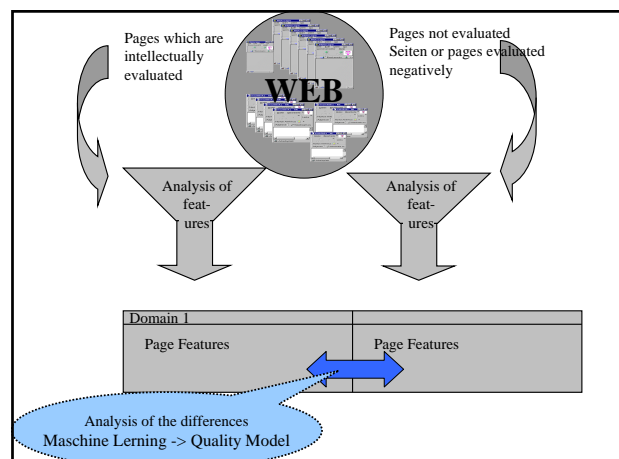


Automatic  
Quality Assessment for Internet  
Resources

AQUAINT was funded by the  
German Research Foundation (DFG)  
Grant MA 2411/3-1

## AQUAINT

- Perspektive: Quality Information Retrieval
- Quality Basis: Decisions made at Internet-Catalogues (Yahoo)
- Other web pages as contrastive (negativ) pages
- Different pages are used for model development and for evaluation
- Evaluation considers retrieval effectivity and page quality



# Features

- Single Features tell us little or are ambivalent
- Example: age of a page
  - Conference pages from last year?
- ->Complex Quality Model
  - Disadvantage: no transparency

- # Features
- Single Features tell us little or are ambivalent
  - Example: age of a page
    - Conference pages from last year?
  - ->Complex Quality Model
    - Disadvantage: no transparency

## AQUAINT: Features

- Features extracted from HTML Code and DOM
  - Some 110 features
  - Partly from previous research
- Examples for features
  - Graphic vs. Text orientation (Colors, Graphics)
  - Structure and complexity
  - Size of some elements (Tags)
  - Text, Links, Hierarchy Level
  - Balance (e.g. between Links and Text ...)

- ## AQUAINT: Features
- Features extracted from HTML Code and DOM
    - Some 110 features
    - Partly from previous research
  - Examples for features
    - Graphic vs. Text orientation (Colors, Graphics)
    - Structure and complexity
    - Size of some elements (Tags)
    - Text, Links, Hierarchy Level
    - Balance (e.g. between Links and Text ...)

## Features: Design



- Design very important for human quality judgement (Tractinsky 1997, Bouch et al. 2000)
  - Eye is primarily directed to graphic elements (Ollermann et al. 2004)
  - Strong correlation between design und trust (Fogg et al. 2001)

- ## Features: Design
- Design very important for human quality judgement (Tractinsky 1997, Bouch et al. 2000)
    - Eye is primarily directed to graphic elements (Ollermann et al. 2004)
    - Strong correlation between design und trust (Fogg et al. 2001)



# Features: Design

- Antagonism (cf. Bürdek 2000, Fries 2004)

Simplicity	Complexity
Structure	complex figures
Symmetry	cluttered
	overburdened





The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.

- # Features: Design
- Antagonism (cf. Bürdek 2000, Fries 2004)
- |            |                 |
|------------|-----------------|
| Simplicity | Complexity      |
| Structure  | complex figures |
| Symmetry   | cluttered       |
|            | overburdened    |
- 
- 
- The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.

# Features: Design

- Antagonism (cf. Bürdek 2000, Fries 2004)

Simplicity	Complexity
Structure	complex figures
Symmetry	cluttered
	overburdened





The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.

# Features: Design

- Antagonism (cf. Bürdek 2000, Fries 2004)

Simplicity	Complexity
Structure	complex figures
Symmetry	cluttered
	overburdened





The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.

# Features: Design

- Antagonism (cf. Bürdek 2000, Fries 2004)

Simplicity	Complexity
Structure	complex figures
Symmetry	cluttered
	overburdened





The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.

# Features: Design

- Antagonism (cf. Bürdek 2000, Fries 2004)

Simplicity	Complexity
Structure	complex figures
Symmetry	cluttered
	overburdened





The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.

# Features: Design

- Antagonism (cf. Bürdek 2000, Fries 2004)

Simplicity	Complexity
Structure	complex figures
Symmetry	cluttered
	overburdened





The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.

# Features: Design

- Antagonism (cf. Bürdek 2000, Fries 2004)

Simplicity	Complexity
Structure	complex figures
Symmetry	cluttered
	overburdened





The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.

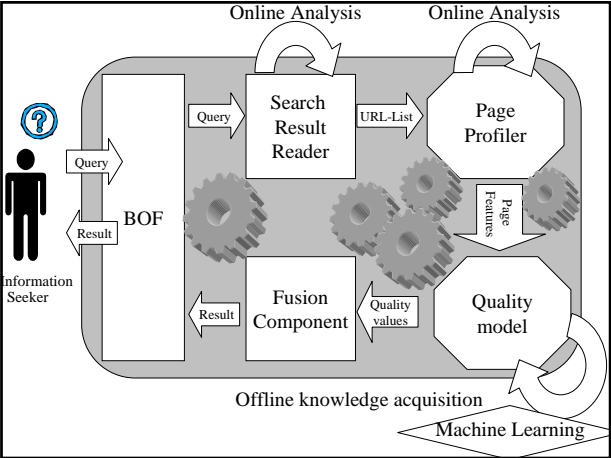
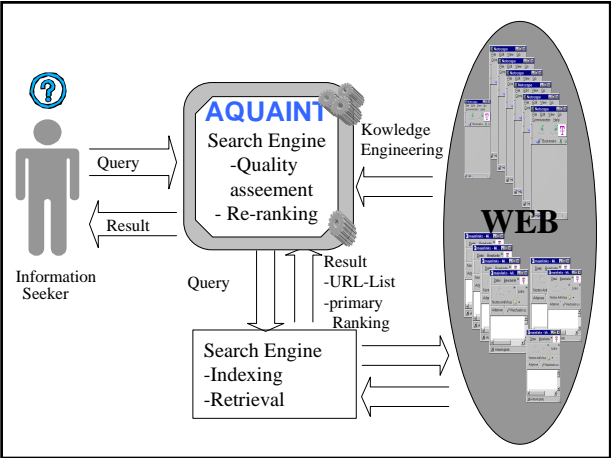
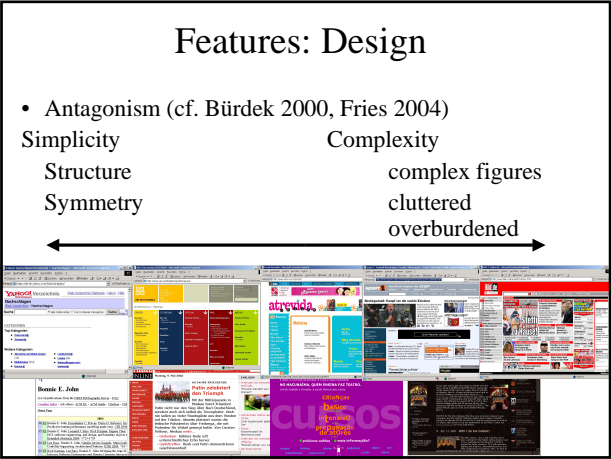
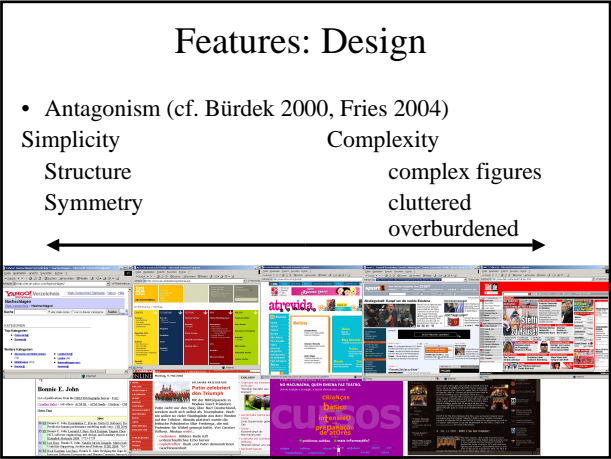
# Features: Design

- Antagonism (cf. Bürdek 2000, Fries 2004)

Simplicity	Complexity
Structure	complex figures
Symmetry	cluttered
	overburdened

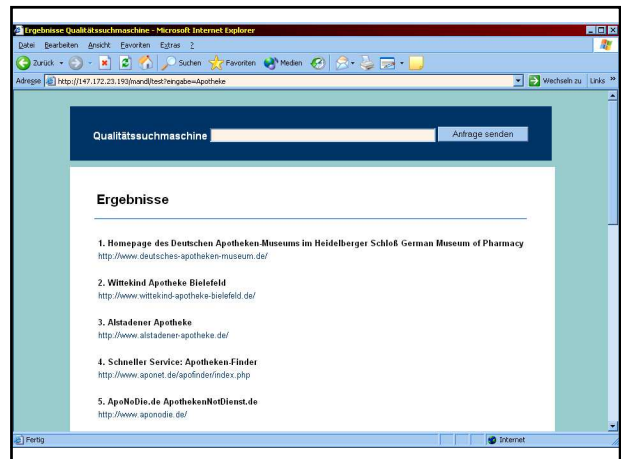


The collage at the bottom shows a variety of web designs. On the left, there are examples of simple, structured, and symmetrical designs, such as a basic navigation menu and a clean layout with a central image. On the right, there are examples of complex, cluttered, and overburdened designs, featuring multiple columns of text, numerous small images, and a dense arrangement of elements. The designs range from minimalist to highly detailed and busy.



## Quality Model

- Current model
  - some 15.000 pages from Yahoo - Health
  - some 15.000 pages from Search engines
  - some 10.000 intellektually found Spam (Source: Lycos Europe)
- Linear Regression Model



## Evaluation



## Evaluation: Subjektivität of Quality Judgements

- “The quality of a web site inherently is a matter of human judgement” (Amtento et al. 2000:296)
- “In fact, for a website there can be as many views of its quality as there are usages” (Brajnik 2001:2)
- “Many kinds of human judgement are intrinsically inconsistent ” (Mizzaro 1997:814)

## Evaluation

- Searches in Domain Health
- Grading of results pages by test users
  - According to relevance and
  - Quality
- 20 test users with 10 queries each
  - Log-File
  - Notes of test administrators

## Evaluation: Subjektivität of Quality

- > **Break with Cranfield-Paradigm of Evaluation in Information Retrieval**
  - No transcendent and absolute relevance
  - But individual, subjective quality evaluation in the context
  - Different evaluation strategy as in standard information retrieval evaluation (TREC, CLEF, NTCIR, INEX, ...)

## Evaluation Results AQUAINT: At Ten Documents

Ranking Method	Grade assigned by user	Quality Grading	Relevance Grading
Original Ranking	Grade 1	29	71
	Grade 1 to 2	101	114
	Grade 1 to 3	154	143
Quality Ranking	Grade 1	32	81
	Grade 1 to 2	119	129
	Grade 1 to 3	185	167
Random Ranking	Grade 1	20	49
	Grade 1 to 2	68	81
	Grade 1 to 3	114	109

## Future Work

- Future Quality Models?
  - Probably combinations of link analysis, content analysis as well as presentation analysis
- Web-Design Mining as a sub task of Web Mining
  - e.g. colors (Eibl & Mandl 2005) or structure (Mandl 2003)

*Thanks for your  
Attention*

*I am looking forward  
to the Discussion*