

# **Focused Access to Wikipedia**

Börkur Sigurbjörnsson  
ISLA, Informatics Institute  
University of Amsterdam

# Motivation

- Experiment with Wikipedia as an IR corpus
  - It's large, free, and **FUN!**
- Evaluate focused retrieval for semi-structured documents
  - Other than scientific publications (INEX)
  - Portability of our XML interface (xmlfind)
  - Study information seeking behavior

# Overview

- Motivation
- **Wikipedia**
- Focused Access to Wikipedia
- Evaluation
- Conclusions

# Wikipedia

- A free encyclopedia
  - Anyone can add and edit entries
- Multilingual
  - English: 1,020,861 pages
  - Dutch: 147.854 pages
  - Mongolian: 278 pages
  - ...□□

# Wikipedia as Test Collection

- Wikipedia attractive corpus for IR
  - It's "free"
    - no paperwork, online demo's
  - It's a general corpus
    - non-technical, easy to invent and assess natural topics, find test persons
  - Highly structured and densely linked
    - apply structured retrieval techniques
- Will be used at QA@CLEF and INEX

# Overview

- Motivation
- Wikipedia
- **Focused Access to Wikipedia**
- Evaluation
- Conclusions

# Focused Access

- What?
  - Access to relevant information within relevant pages
- Why?
  - More appropriate access if information need is specific
- How?
  - Retrieve sections instead of documents

# Extract sections

## Economics

- Definitions of economics
  - Wealth definition
  - Welfare definition
  - Scarcity definition
- Areas of study in economics
- Economic assumptions
  - Supply and demand
  - Price
  - Scarcity
  - Marginalism
  - Value
- Economic language and reasoning
- Development of economic thought
- Schools of economic thought
  - Modern 'mainstream' economics
  - Neoclassical economics
  - Post-Keynesian economics

The screenshot shows the Wikipedia page for 'Economics' as of March 2006. Red boxes highlight the following sections:

- Economic definitions:** A paragraph defining economics as the study of production, distribution, and consumption of goods and services, and its division into microeconomics and macroeconomics.
- Definitions of economics:** A section discussing the history of the term 'economics' and its evolution from 'political economy'.
- Wealth definition:** A section discussing the definition of wealth as the stock of useful things, and its relationship to production and consumption.
- Welfare definition:** A section discussing the definition of welfare as the attainment and use of material requisites of well-being, and its relationship to production and consumption.
- Scarcity definition:** A section discussing the definition of scarcity as the condition in which the demand for a good exceeds the supply, and its relationship to production and consumption.

# Retrieval

- Index each section as separate unit
  - Section retrieval
  - Cluster by article
- 
- So, it resembles a standard search engine
  - But with direct access to individuals sections

# Interface

- Each article
  - Links to sections
  - Section level snippets
- Let's have a look...
  - <http://berk.science.uva.nl:8080/wikiii/result.php>

# Overview

- Motivation
- Wikipedia
- Focused Access to Wikipedia
- **Evaluation**
- Conclusions

# Evaluation Setup

- Interactive experiment
  - Performed as part of student project
- Systems
  - Baseline system
  - Focused system
- Subjects solved 2 simulated work-tasks
  - Using baseline system
  - Using focused system

# Experiment setup

- Experiment setup
  - Pre-experiment questionnaire
  - Simulated work task I
    - Post-task questionnaire
  - Simulated work task II
    - Post-task questionnaire
  - Post-experiment questionnaire
- System rotation
  - Task I: Focused    Task II: Baseline
  - Task II: Baseline    Task I: Focused

# Research Questions

- Satisfaction
  - Do users appreciate focused access?
- Usefulness
  - Is focused access useful?
- User interaction
  - Page vs. section clicks
  - Fetch vs. Browse

# Satisfaction

- How satisfied are you with the answers given by the system?
  - 5-point scale:
    - very dissatisfied ... very satisfied

	<b>Task I</b>	<b>Task II</b>	<b>Overall</b>
Baseline	<b>4.17</b>	3.00	3.58
Focused	3.67	<b>3.67</b>	<b>3.67</b>

# Effort

- The answers to the task-questions were in this system ...
  - 5-point scale
    - difficult to find ... easy to find

	<b>Task I</b>	<b>Task II</b>	<b>Overall</b>
Baseline	<b>3.17</b>	2.83	3.00
Focused	2.67	<b>3.50</b>	<b>3.08</b>

# Page views

- Number of Wikipedia pages viewed while solving work task

	Task I	Task II	Overall
Baseline	<b>19.2</b>	16.3	17.8
Focused	15.5	<b>26.0</b>	<b>20.8</b>

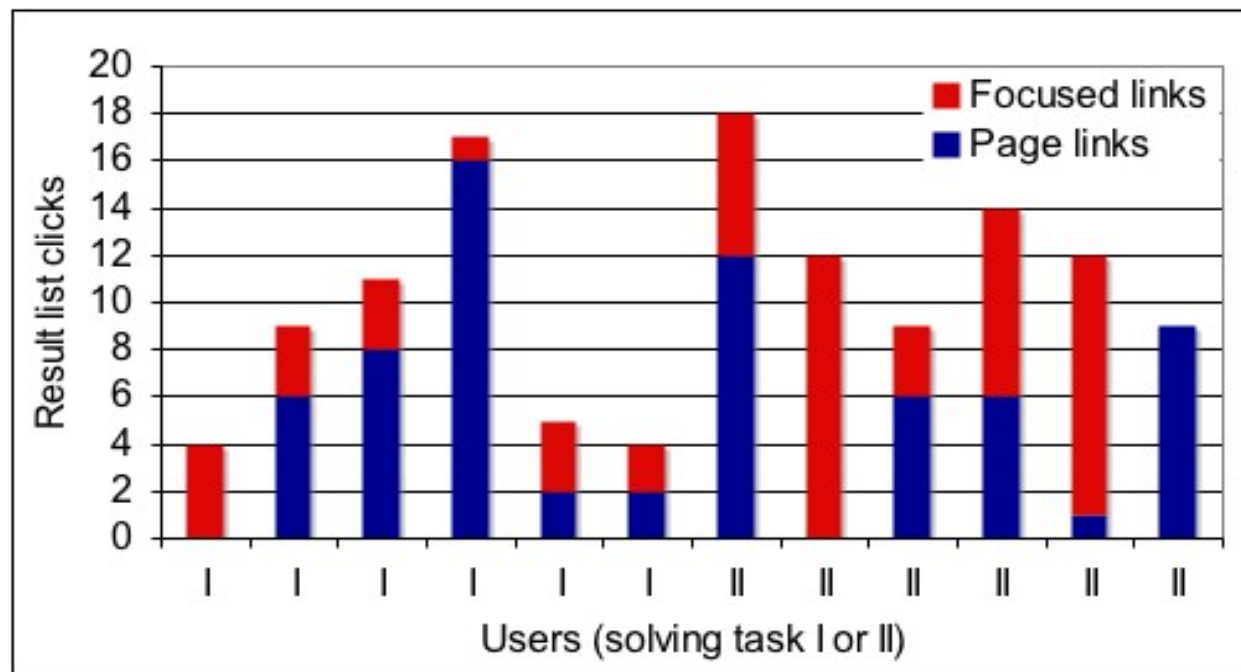
# Time to solve task

- Time needed to complete task

	<b>Task I</b>	<b>Task II</b>	<b>Overall</b>
Baseline	<b>31.2</b>	<b>27.0</b>	<b>29.1</b>
Focused	23.3	22.5	22.9

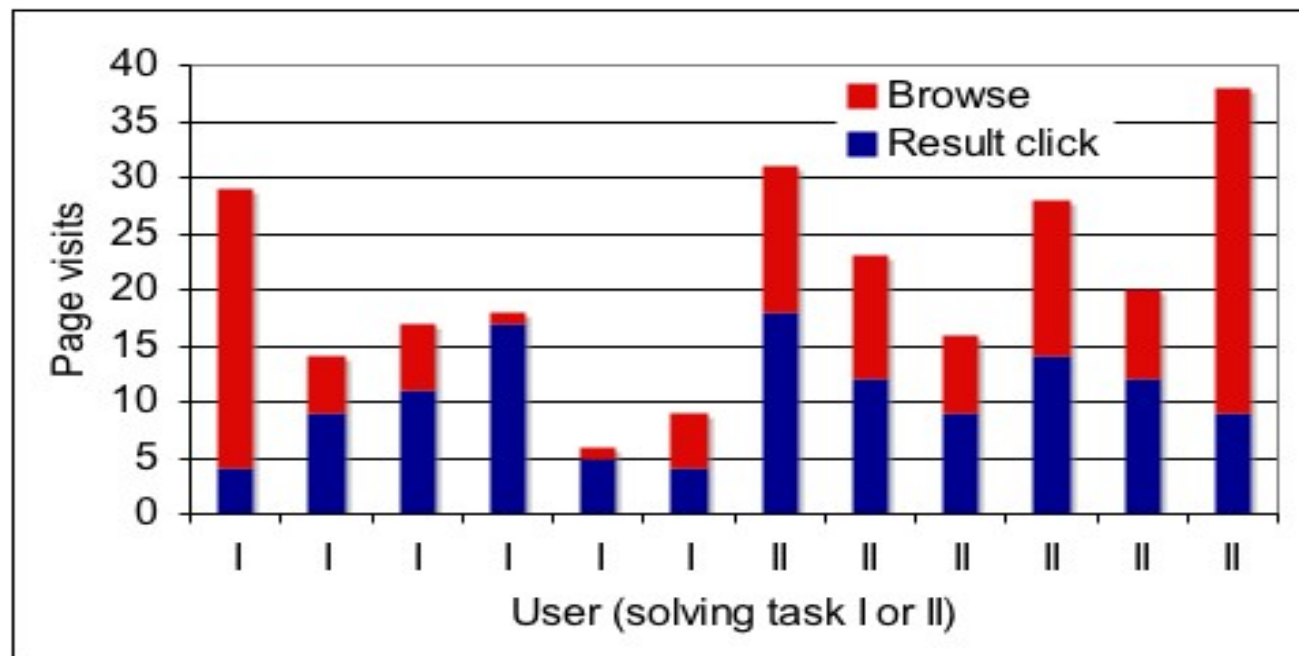
# Page-links vs. Focused-links

	Task I	Task II	Overall
Page-links	<b>68%</b>	46%	<b>55%</b>
Focused-links	32%	<b>54%</b>	45%



# Result clicks vs. Browse

	Task I	Task II	Overall
Result click	<b>61%</b>	50%	<b>55%</b>
Browse	39%	50%	45%



# Overview

- Motivation
- Wikipedia
- Focused Access to Wikipedia
- Evaluation
- **Conclusions**

# Wrap Up

- Wikipedia attractive corpus for IR
  - Here used to evaluate focused retrieval
- Main findings:
  - Focused engine helps users solve search tasks faster
  - Dense link structure: users access pages as often by searching and browsing
- Interesting interaction between searching and browsing due to structure