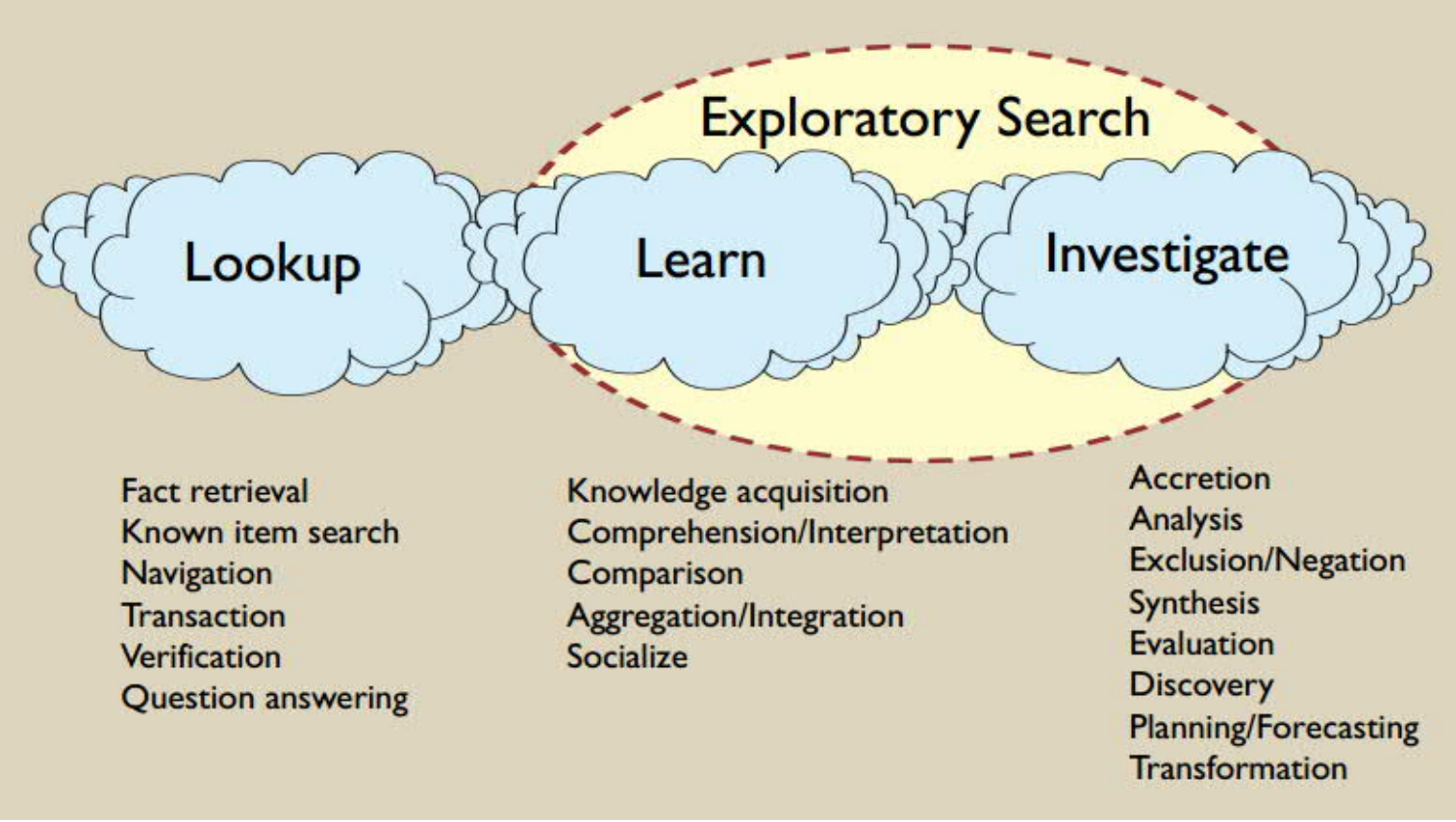
Do not search, explore: Supporting scientific literature exploration

Dorota Glowacka
University of Helsinki

Exploratory Search



G. Marchionini. Exploratory search: from finding to understanding. Comm. ACM, 49(4):41–46, 2006

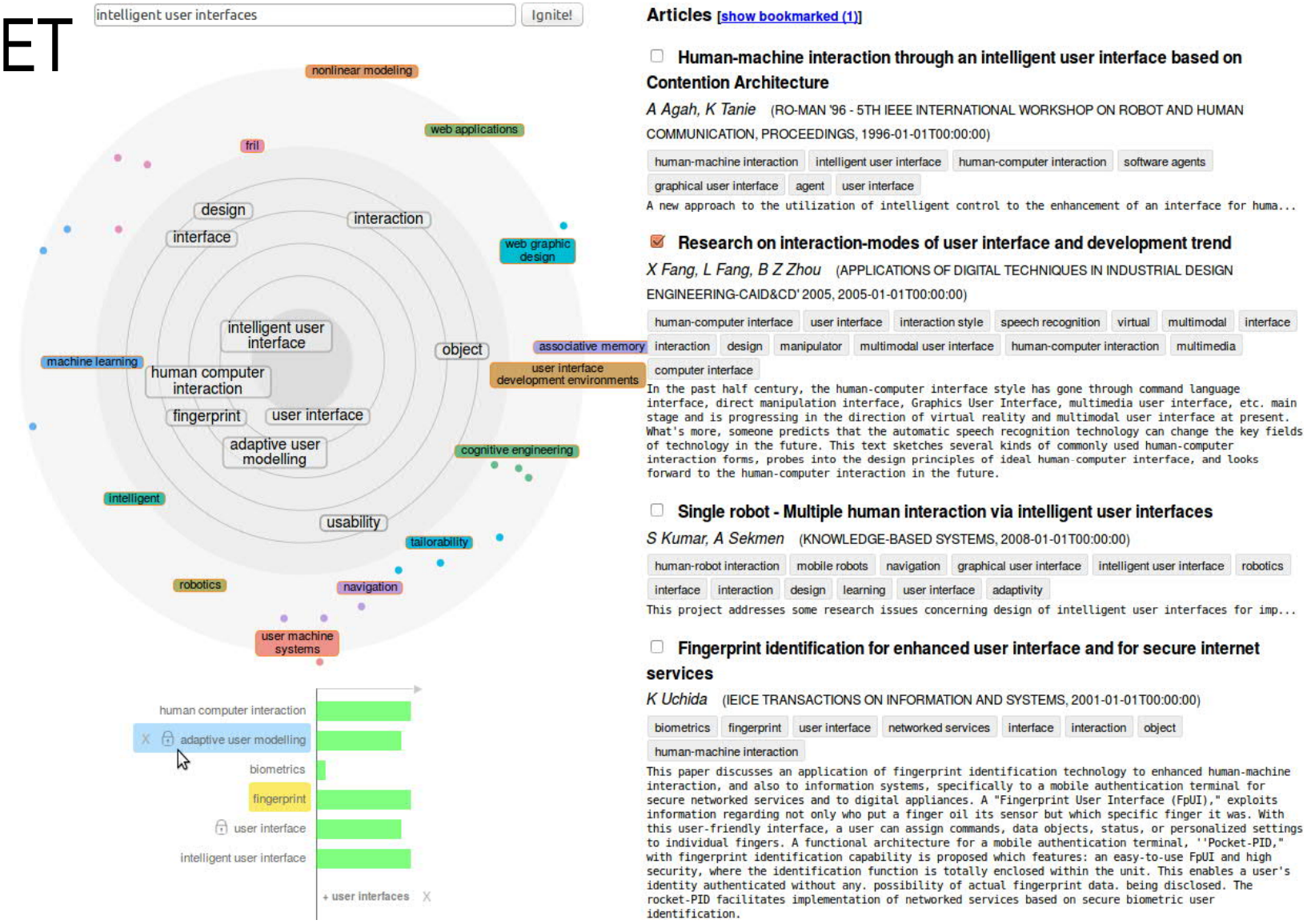
Core research questions

- How to support the user in navigating the information space?
- Search goals can be poorly defined. How to understand the user’s search intent?
- Information needs are dynamic. How to adapt with the user?

Approach

- Build practical information retrieval systems for exploratory search using reinforcement learning
- Draw insights from user studies to understand user behaviour and fine tune the system
- Draw insights from user behaviour to improve user modelling (and iterate …)

SCINET



Articles [show bookmarked (1)]

Human-machine interaction through an intelligent user interface based on Contention Architecture

A Agah, K Tanie (RO-MAN '96 - 5TH IEEE INTERNATIONAL WORKSHOP ON ROBOT AND HUMAN COMMUNICATION, PROCEEDINGS, 1996-01-01T00:00:00)

human-machine interaction | intelligent user interface | human-computer interaction | software agents | graphical user interface | agent | user interface

A new approach to the utilization of intelligent control to the enhancement of an interface for huma...

Research on interaction-modes of user interface and development trend

X Fang, L Fang, B Z Zhou (APPLICATIONS OF DIGITAL TECHNIQUES IN INDUSTRIAL DESIGN ENGINEERING-CAID&CD' 2005, 2005-01-01T00:00:00)

human-computer interface | user interface | interaction style | speech recognition | virtual | multimodal | interface | interaction | design | manipulator | multimodal user interface | human-computer interaction | multimedia | computer interface

In the past half century, the human-computer interface style has gone through command language interface, direct manipulation interface, Graphics User Interface, multimedia user interface, etc. main stage and is progressing in the direction of virtual reality and multimodal user interface at present. What's more, someone predicts that the automatic speech recognition technology can change the key fields of technology in the future. This text sketches several kinds of commonly used human-computer interaction forms, probes into the design principles of ideal human-computer interface, and looks forward to the human-computer interaction in the future.

Single robot - Multiple human interaction via intelligent user interfaces

S Kumar, A Sekmen (KNOWLEDGE-BASED SYSTEMS, 2008-01-01T00:00:00)

human-robot interaction | mobile robots | navigation | graphical user interface | intelligent user interface | robotics | interface | interaction | design | learning | user interface | adaptivity

This project addresses some research issues concerning design of intelligent user interfaces for imp...

Fingerprint identification for enhanced user interface and for secure internet services

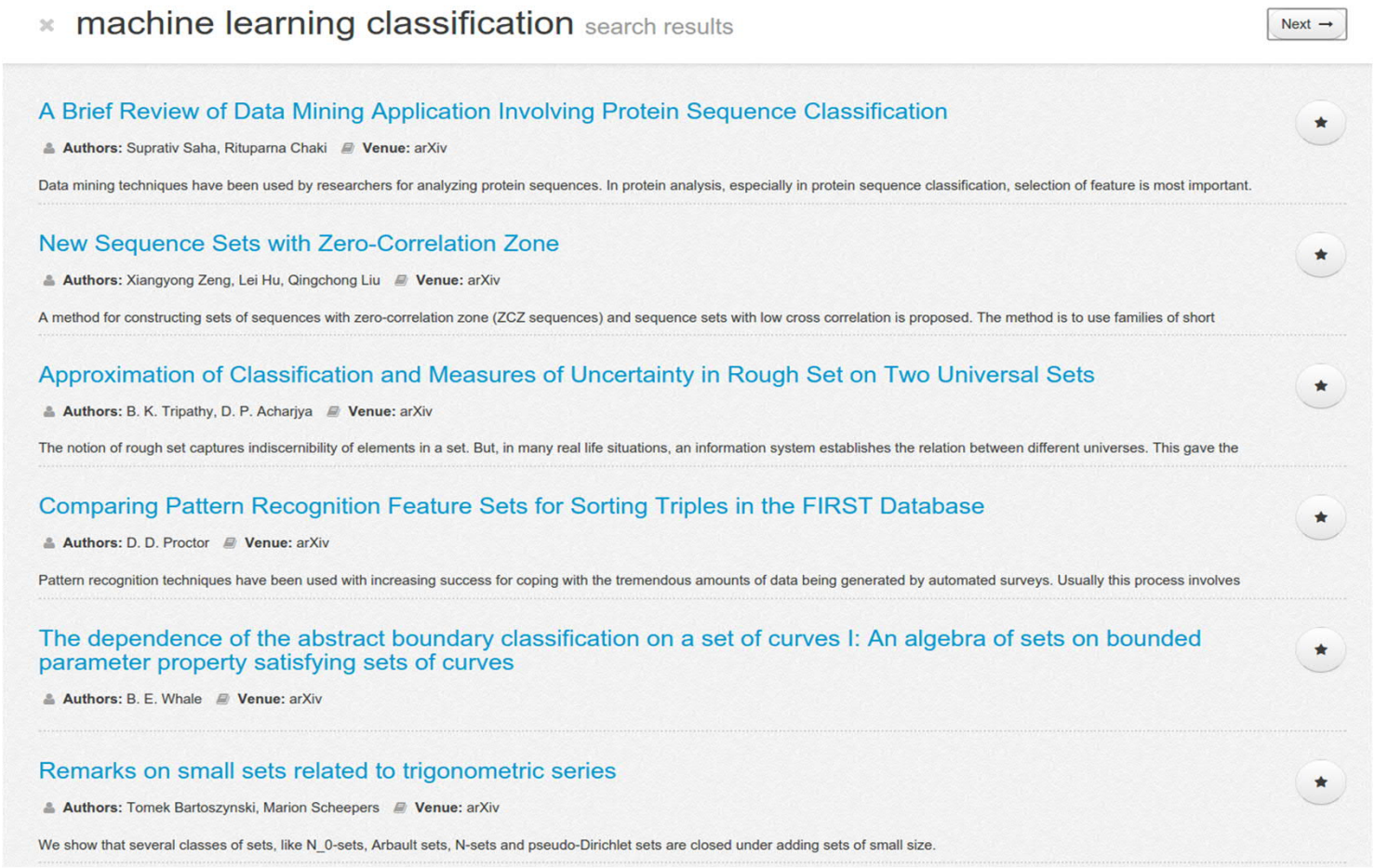
K Uchida (IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS, 2001-01-01T00:00:00)

biometrics | fingerprint | user interface | networked services | interface | interaction | object | human-machine interaction

This paper discusses an application of fingerprint identification technology to enhanced human-machine interaction, and also to information systems, specifically to a mobile authentication terminal for secure networked services and to digital appliances. A "Fingerprint User Interface (FpUI)," exploits information regarding not only who put a finger oil its sensor but which specific finger it was. With this user-friendly interface, a user can assign commands, data objects, status, or personalized settings to individual fingers. A functional architecture for a mobile authentication terminal, ''Pocket-PID,'' with fingerprint identification capability is proposed which features: an easy-to-use FpUI and high security, where the identification function is totally enclosed within the unit. This enables a user's identity authenticated without any. possibility of actual fingerprint data. being disclosed. The rocket-PID facilitates implementation of networked services based on secure biometric user identification.

Glowacka et al. IUI 2013, Kangasraasio et al. IUI 2015, Kangasraasio et al. UMAP 2016

PULP



Athukorala et al. CIKM15, Athukorala et al. IUI16, Medlar et al. SIGIR16, Medlar et al. IUI17

Why Reinforcement Learning?

- Generic framework for building user model over a search session
- Allows a system to balance exploration and exploitation
  - Users are not trapped in a local context bubble based on the initial query
  - Users exposed to more diverse set of data
  - Harder problem than expected ….

Balancing Exploration/Exploitation

- Implicit feedback (Athukorala et al. CIKM15)
  - Model the relationship between exploration rate and number of relevant documents
  - Offline – one exploration rate for population
- Explicit feedback (Medlar et al. IUI17)
  - Model the relationship between exploration rate and user experience
  - Online – personalised exploration rate based on interaction data

Balancing Exploration/Exploitation

- Too low – stuck in context bubble, too high – results appear random ... but this is subjective
- We want to model the relationship between the exploration rate and user experience
- Novel approach based on interval regression

Study Design

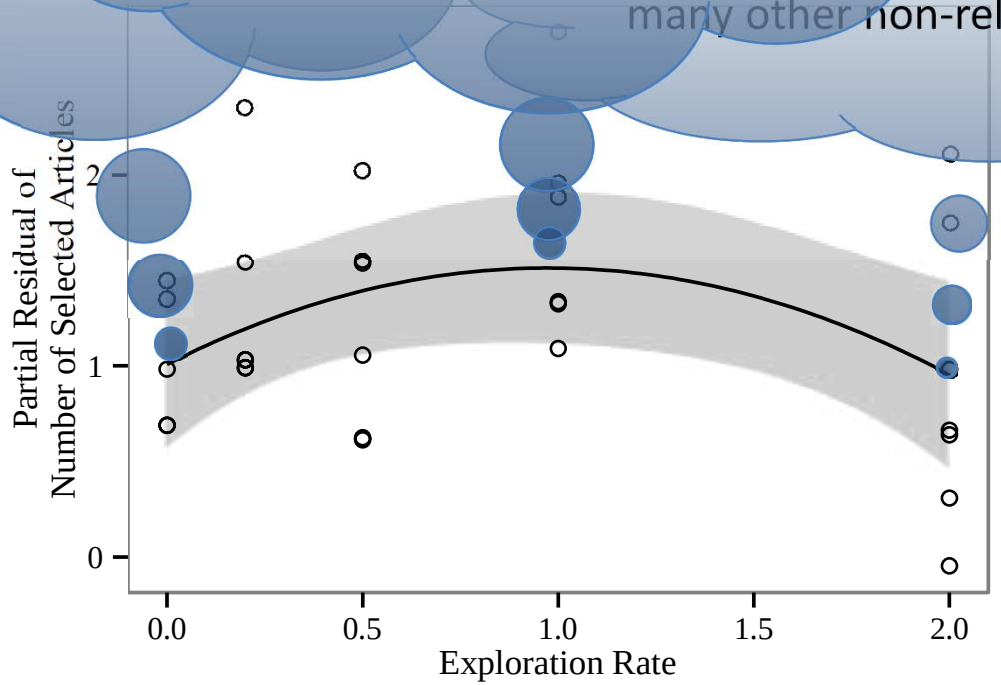
- Simulations: exploration rates to show different numbers of “exploratory” documents
- User study: MSc/PhD researchers in Machine Learning, 5 ML queries using different exploration rates
- Analysis: modelling combined with qualitative analysis of user performance data

Results

"Results are not quite satisfactory and about what specific definitions. I did not get [any] understanding of what topics are in this area"

"I went over several iterations. Results started getting way better [over iterations] and overall I am very satisfied"

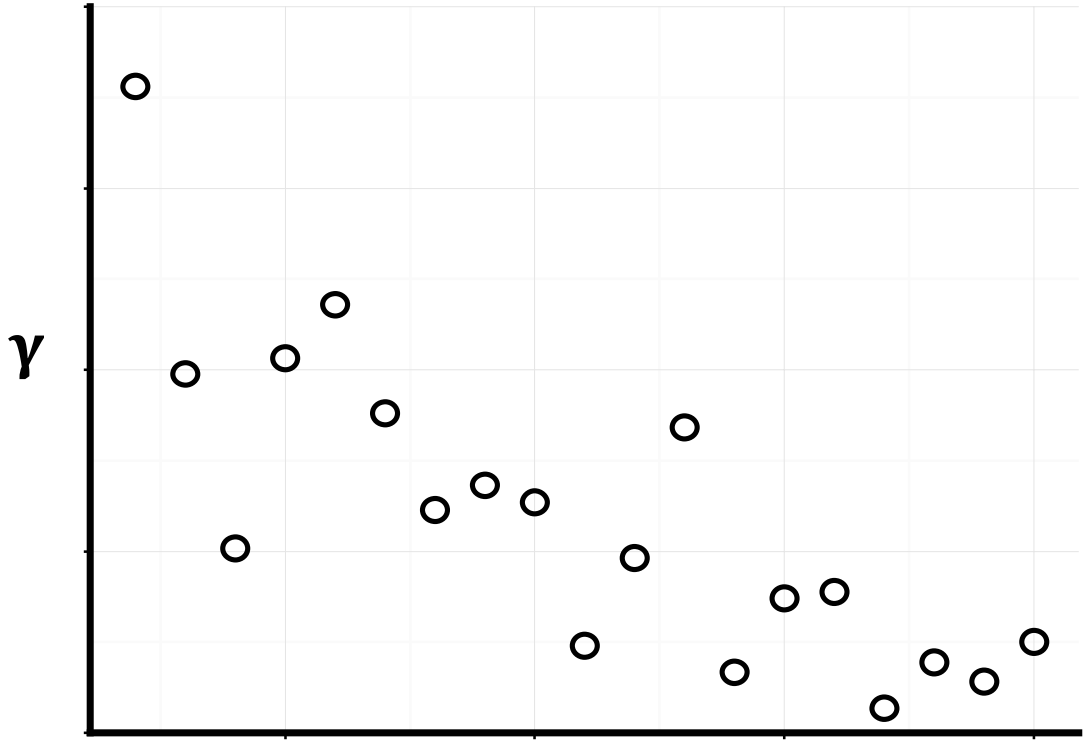
"[results are] too scattered and many other non-related papers"



Can we build a regression model to adjust the exploration rate per user per session?

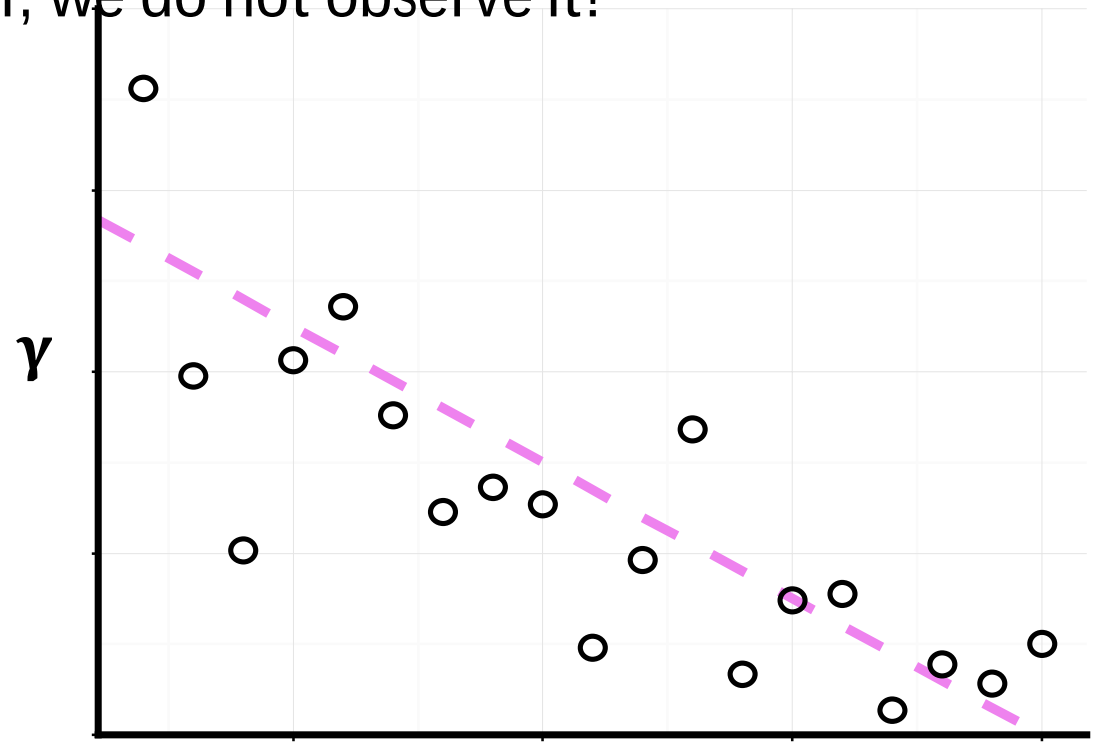
Setting Exploration/Exploitation with Explicit Feedback

Ordinary regression fits a model based on linear relationships between response and explanatory variables...



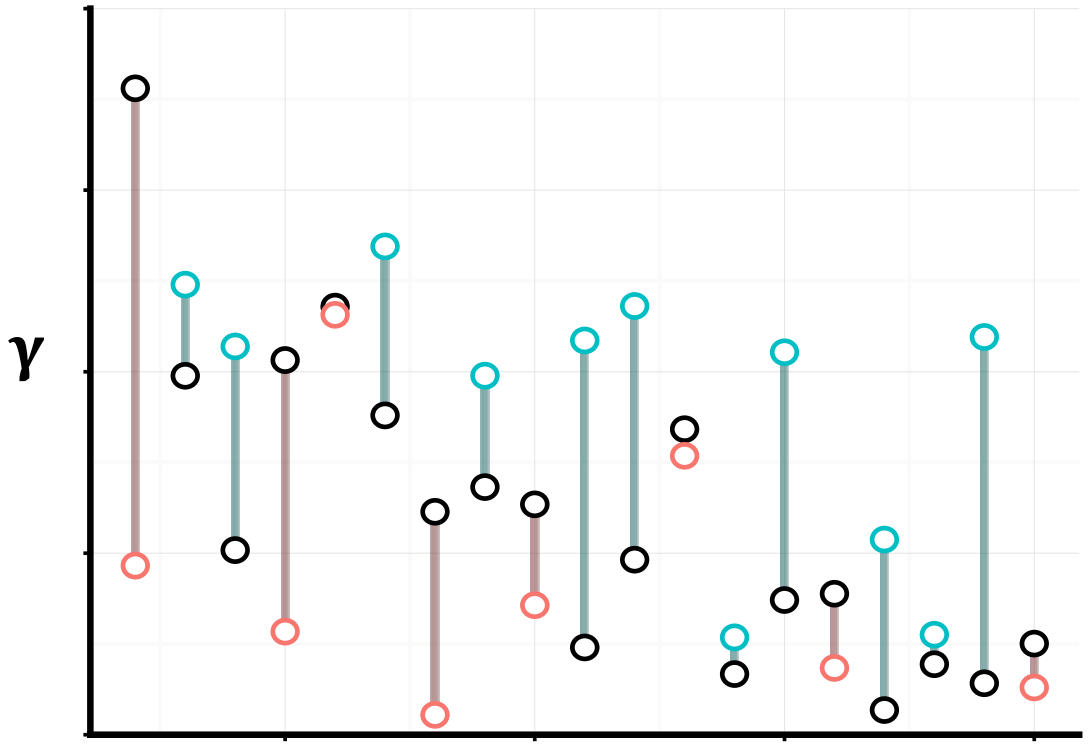
Setting Exploration/Exploitation with Explicit Feedback

Ordinary regression fits a model based on linear relationships between response and explanatory variables, but $\boldsymbol{\gamma}$ is a parameter, we do not observe it!



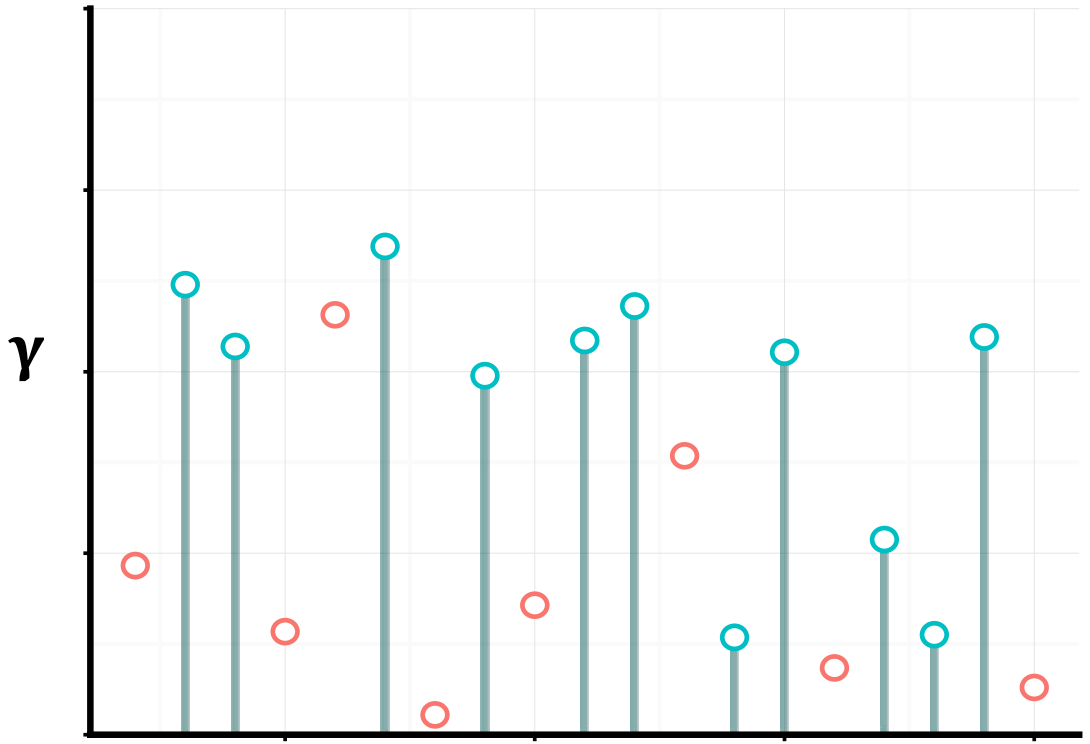
Setting Exploration/Exploitation with Explicit Feedback

Run experiments with random γ and collect user feedback. For some users, γ was too low, for others γ was too high



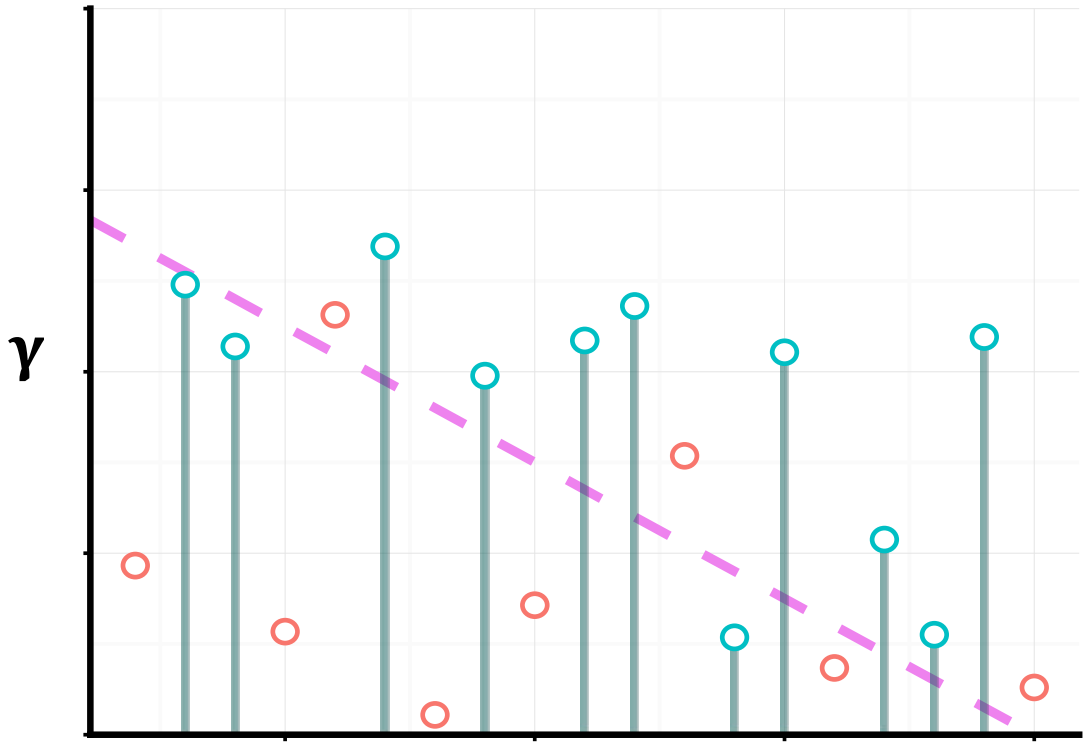
Setting Exploration/Exploitation with Explicit Feedback

Don't know "true" γ, but if too high, "true" γ will be in interval $[0, \gamma]$ and if too low, "true" γ will be in interval $[\gamma, \infty]$



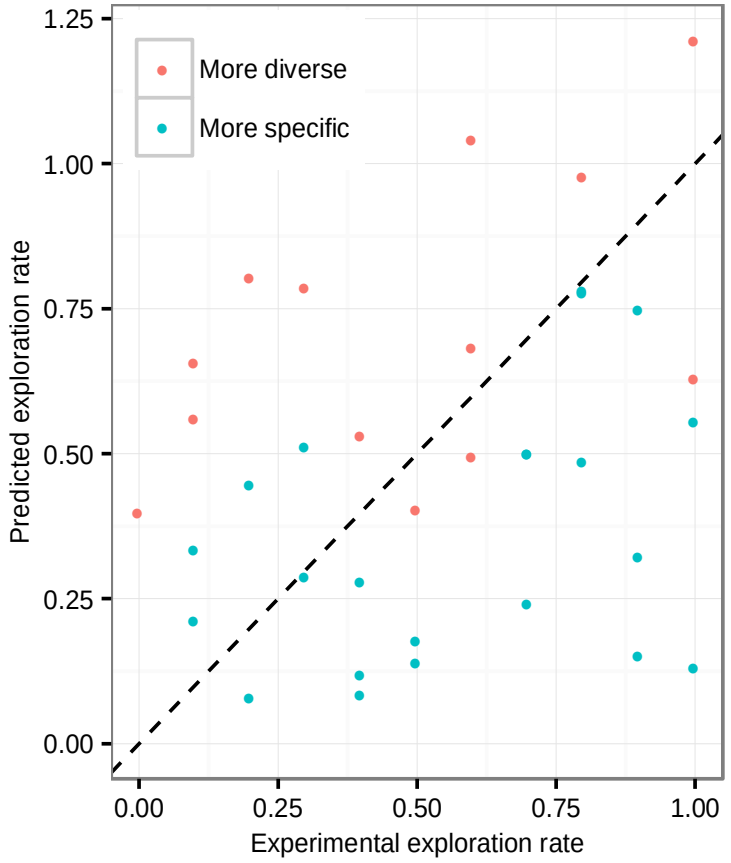
Setting Exploration/Exploitation with Explicit Feedback

Fit model over censored intervals and then make predictions as normal!



Balancing Exploration/Exploitation with User Experience

- Users state document diversity preferences relative to random exploration rate
- Three user interaction variables (knowledge level, interface time, clicked documents)
- Prediction consistent with user feedback in 80% of cases (Medlar et al. IUI17)



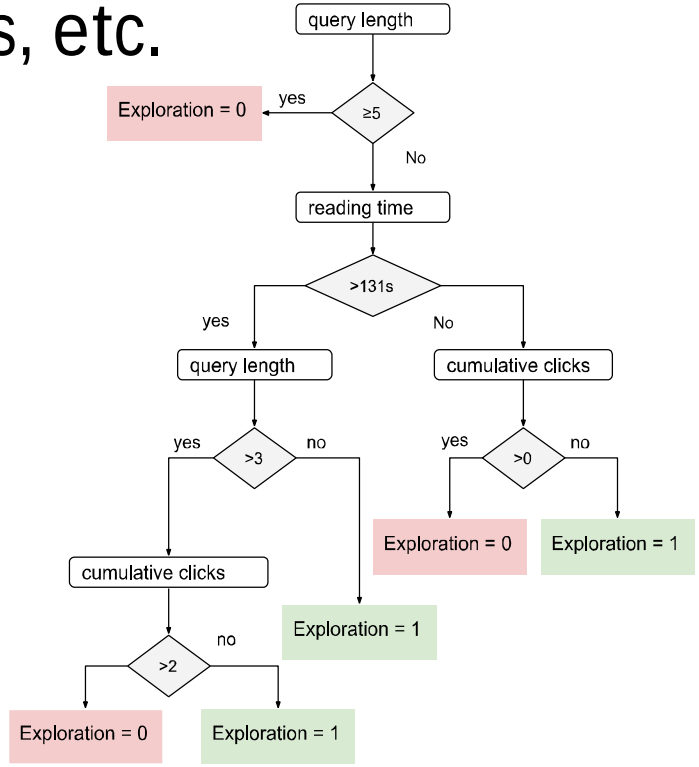
Analysing User Behaviour in Exploratory Search

- Building new IR systems will only get us so far…
- Need to develop a better understanding of user behavior to improve user modelling
  - Is exploratory search different from lookup search?
  - How reliable is relevance feedback (assume users know what they want and are consistent)

User Behaviour in Exploratory Search vs. Lookup Search

- People behave differently when performing exploratory and lookup searches (Athukorala et al. JASIST16): longer queries, more scrolling, longer search sessions, more clicks, etc.

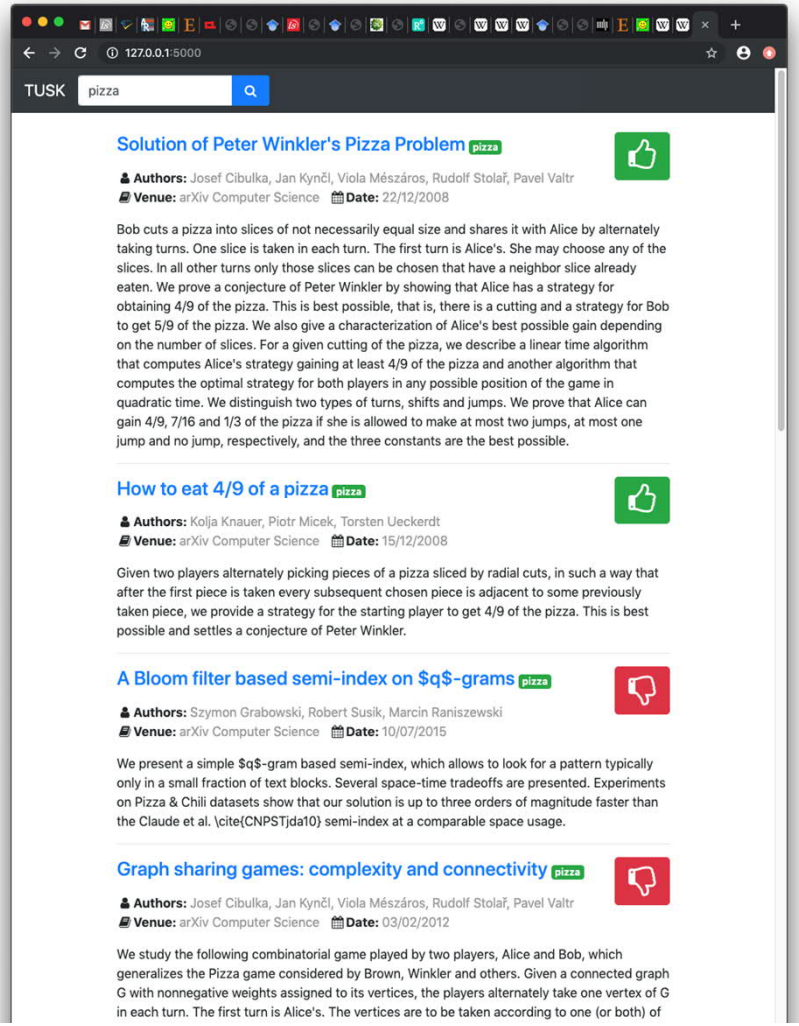
- Differences are predictive: simple classifier predicts exploratory/lookup with 81% accuracy (Athukorala et al. IUI16).



SCIENTIFIC LITERATURE AND ARTICLE STRUCTURE

Motivation

- We perform exploratory search user studies on scientific literature
- Full-text retrieval has higher recall (and lower precision) over searching bibliographic records (title, abstract, etc.)
- Could using abstracts for retrieval impact experimental results?



Abstracts vs. full-text

Holes in the Outline: Subject-dependent Abstract Quality and its Implications for Scientific Literature Search

Chien-yu Huang and Arlene Casey
University of Edinburgh
s1771650@sms.ed.ac.uk; A.J.Casey@sms.ed.ac.uk

Dorota Glowacka and Alan Medlar
University of Helsinki
glowacka@cs.helsinki.fi; alan.j.medlar@helsinki.fi

ABSTRACT

Scientific literature search engines typically index abstracts instead of the full-text of publications. The expectation is that the abstract provides a comprehensive summary of the article, enumerating key points for the reader to assess whether their information needs could be satisfied by reading the full-text. Furthermore, from a practical standpoint, obtaining the full-text is more complicated due to licensing issues, in the case of commercial publishers, and resource limitations of public repositories and pre-print servers.

In this article, we use topic modelling to represent content in abstracts and full-text articles. Using Computer Science as a case study, we demonstrate that how well the abstract summarises the full-text is subfield-dependent. Indeed, we show that abstract representativeness has a direct impact on retrieval performance, with poorer abstracts leading to degraded performance. Finally, we present evidence that how well an abstract represents the full-text of an article is not random, but is a consequence of style and writing conventions in different subdisciplines and can be used to infer an "evolutionary" tree of subfields within Computer Science.

CCS CONCEPTS

• **Information systems** → *Document structure; Document collection models;*

KEYWORDS

scientific literature search, topic models, term taxonomy

ACM Reference Format:

Chien-yu Huang and Arlene Casey and Dorota Glowacka and Alan Medlar. 2019. Holes in the Outline: Subject-dependent Abstract Quality, and its Implications for Scientific Literature Search. *In 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298953

1 INTRODUCTION

In scientific literature search abstracts are frequently used as the primary source for indexing and as snippet text that is displayed with search results. The reasons for this are twofold. First, readers assume that abstracts convey sufficient information about the paper

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298953

ACM DL DIGITAL LIBRARY — University of Helsinki including Helsinki University Central Hospital — SIGN IN SIGN UP SEARCH

Check out the beta version of the *next* ACM DL

Holes in the Outline: Subject-dependent Abstract Quality and its Implications for Scientific Literature Search

Full Text: PDF

Authors: Chien-yu Huang, University of Edinburgh, Edinburgh, United Kingdom; Arlene Casey, University of Edinburgh, Edinburgh, United Kingdom; Dorota Glowacka, University of Helsinki, Helsinki, Finland; Alan Medlar, University of Helsinki, Helsinki, Finland

2019 Article

Bibliometrics
· Citation Count: 0
· Downloads (cumulative): 56
· Downloads (12 Months): 56
· Downloads (6 Weeks): 4

Published in:
· Proceeding
CHIIR '19 Proceedings of the 2019 Conference on Human Information Interaction and Retrieval
Pages 289-293
Glasgow, Scotland UK — March 10 - 14, 2019
ACM New York, NY, USA ©2019
table of contents ISBN: 978-1-4503-6025-8 doi>10.1145/3295750.3298953

Tools and Resources
Request Permissions
TOC Service: Email RSS
Save to Binder
Export Formats: BibTeX EndNote ACM Ref
Upcoming Conference: AIRS '19
Share:
Author Tags

Contact Us | Switch to single page view (no tabs)

Abstract | Authors | References | Cited By | Index Terms | Publication | Reviews | Comments | Table of Contents

Scientific literature search engines typically index abstracts instead of the full-text of publications. The expectation is that the abstract provides a comprehensive summary of the article, enumerating key points for the reader to assess whether their information needs could be satisfied by reading the full-text. Furthermore, from a practical standpoint, obtaining the full-text is more complicated due to licensing issues, in the case of commercial publishers, and resource limitations of public repositories and pre-print servers.

In this article, we use topic modelling to represent content in abstracts and full-text articles. Using Computer Science as a case study, we demonstrate that how well the abstract summarises the full-text is subfield-dependent. Indeed, we show that abstract representativeness has a direct impact on retrieval performance, with poorer abstracts leading to degraded performance. Finally, we present evidence that how well an abstract represents the full-text of an article is not random, but is a consequence of style and writing conventions in different subdisciplines and can be used to infer an "evolutionary" tree of subfields within Computer Science.

- Bioinformatics: article sections (e.g. results) besides abstract provides better representations of certain biological concepts
- Medicine: clinical decisions based solely on abstracts results in worse patient outcomes
- General: well-known differences between established/emerging fields, theoretical/applied fields, individual fields tend to have own style/expectations

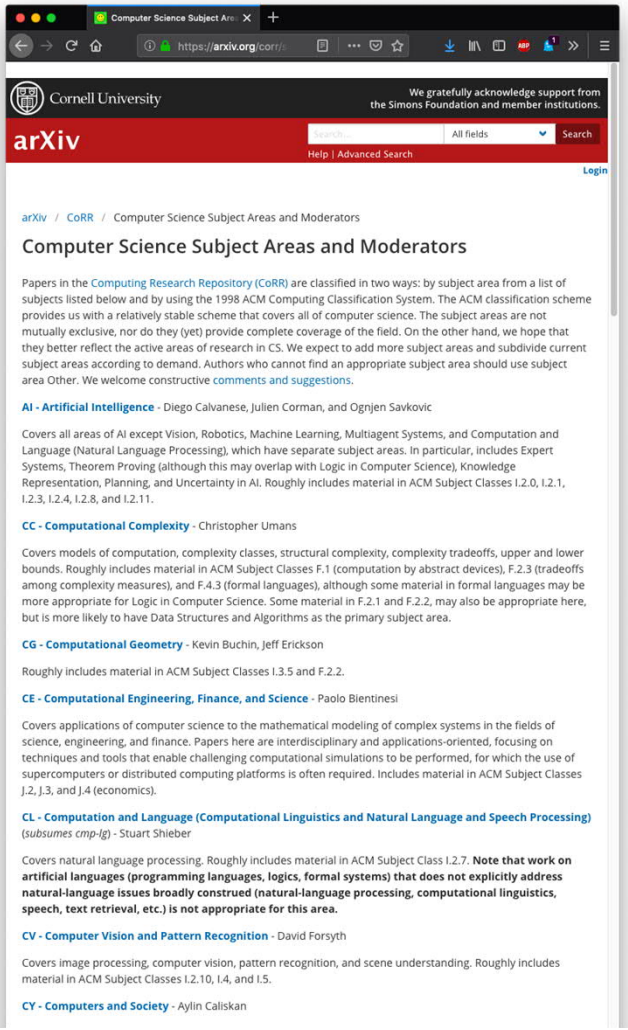
Research Questions

- RQ1: How well do abstracts represent the full-text of a paper in different CS subfields?

- RQ2: If there are differences between subfields, could this impact (perceived) retrieval performance?

Data preprocessing

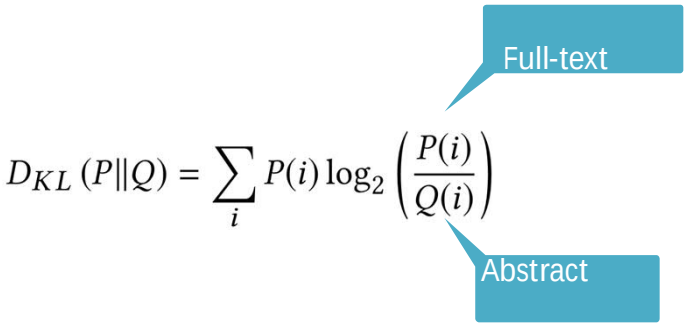
- 35,137 CS papers from arXiv (2007-mid 2015)
  - 23% papers associated with > 1 category (40 possible author-assigned categories)
  - Extracted 6.7 sections per article (SD = 2.7)
- Classify sections as abstract, introduction, background, related work, methods, results, discussion, conclusions and back matter
  - Classified 53% of sections based on headings that occurred at least twice (35% of headings were unique)
  - Classified 3.3 sections per article (SD = 1.2)



Representation

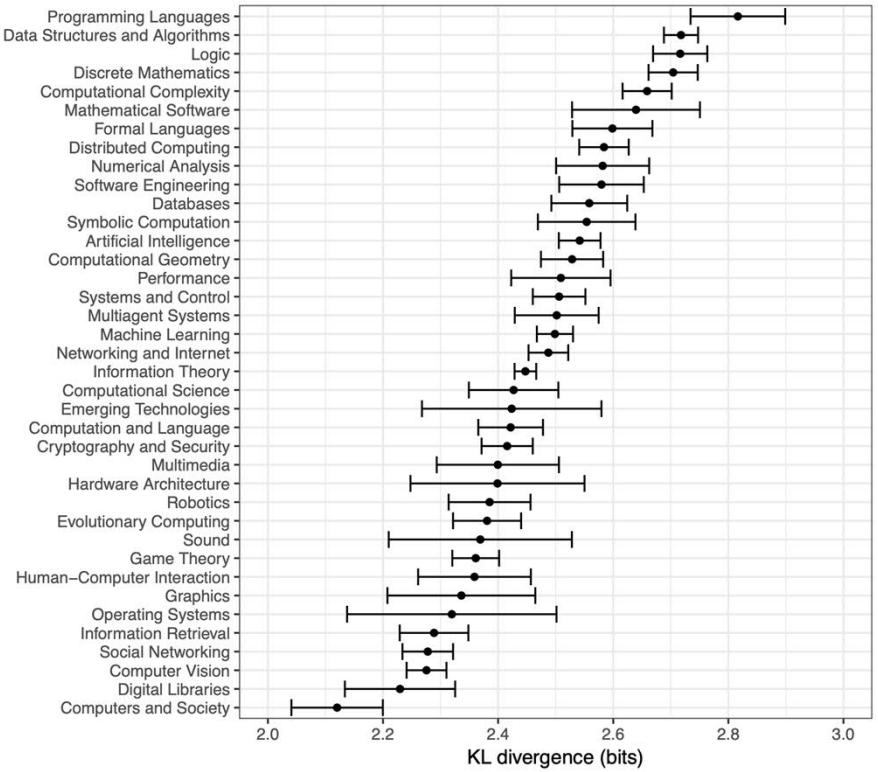
- Full-text and sections represented using probabilistic topic models
  - Topic model inferred from full-text + used to predict individual sections (100 topics)
  - Multi-sections merged using element-wise summarisation and normalised

- Representativeness metric
  - KL divergence: two discrete probability distributions, P and Q, "how much information is lost when Q is used to approximate P"



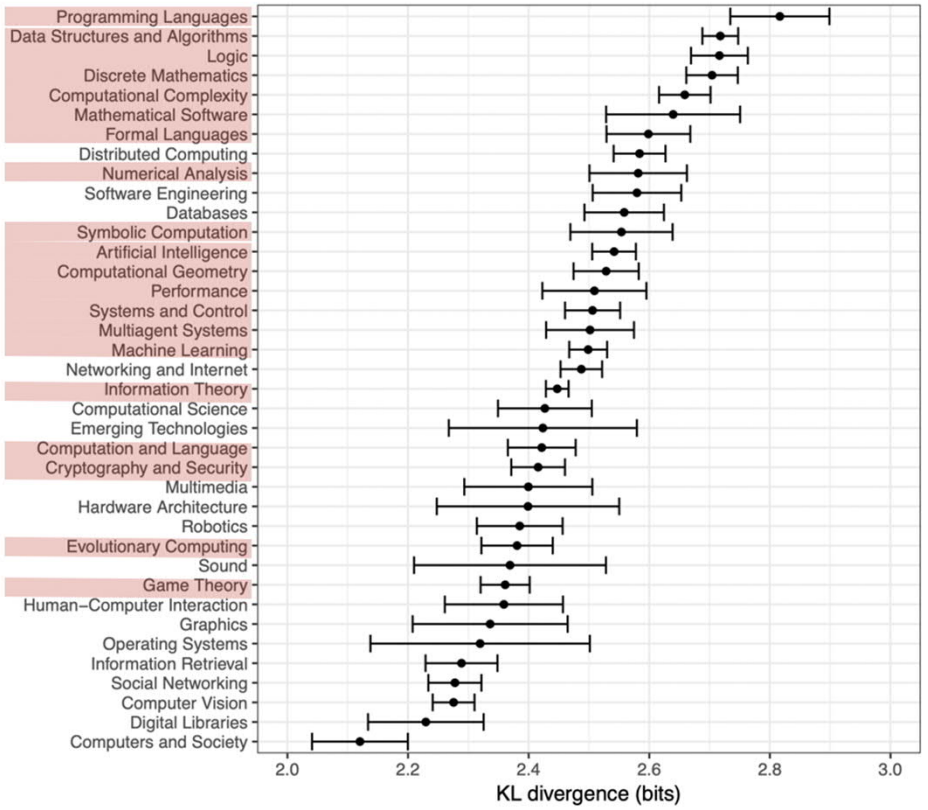
Abstract representativeness is subfield-specific

- The degree to which abstracts represent the full-text is subfield-specific
- Higher mean KL divergence (less representative abstract) appear to be more theoretical...
- Lower mean KL divergence (more representative abstract) appear to be more applied...
- Theoretical abstracts tend to be shorter, but KL divergence is not correlated with abstract length ($R^2 = 0.003$, $p < 2.2 \times 10^{-16}$)



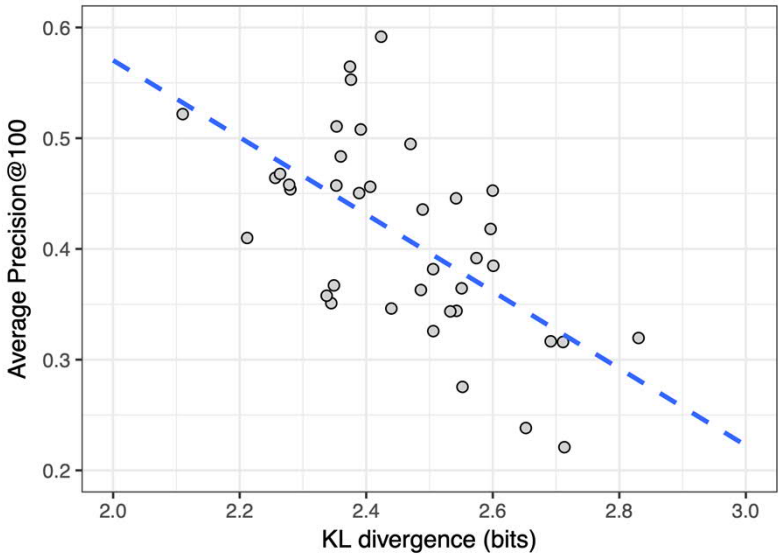
Abstract representativeness is subfield-specific

- The degree to which abstracts represent the full-text is subfield-specific
- Higher mean KL divergence (less representative abstract) appear to be more theoretical...
- Lower mean KL divergence (more representative abstract) appear to be more applied...
- Theoretical abstracts tend to be shorter, but KL divergence is not correlated with abstract length ($R^2 = 0.003$, $p < 2.2 \times 10^{-16}$)



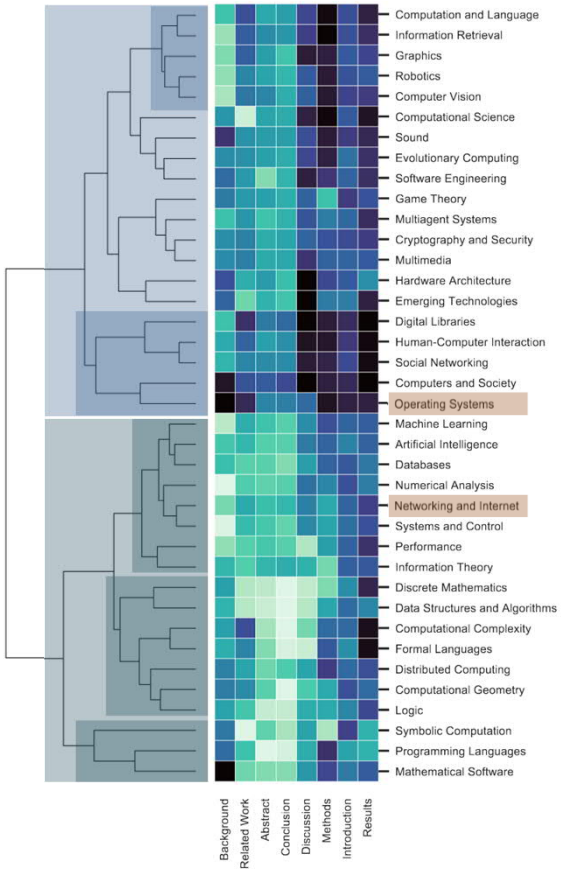
Abstract representativeness is correlated with retrieval performance

- What is the impact on retrieval?
  - Generate queries that disproportionately favour specific subfields (most informative features from multi-class SVM, removed duplicates, manually removed junk)
  - 1,257 queries (33.1 per category, SD = 14.5)
  - Retrieve top-100 results using full-text and abstracts - calculate precision@100
- Precision@100 negatively correlated with KL divergence (R^2 = 0.38, p = 3.65 x 10^{-5})



Section-wise representativeness captures domain structure

- Are these trends random, or related to some underlying structure?
  - Calculate KL divergence between all 8 sections and full-text
  - Do hierarchical clustering (complete-linkage clustering with Euclidean distance)
- Theoretical and applied subtrees, deeper subtrees make sense
- Obvious errors explained by high variance (e.g. Operating Systems) or corpus bias (e.g. Networking and Internet, 23% associated with Information theory as well)



Questions?