Using Annotations in Enterprise Search

Pavel Dmitriev, *Cornell Univ.**

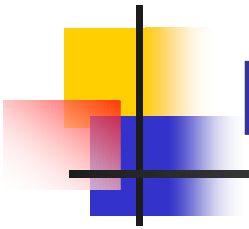
Nadav Eiron, *Google Inc.**

Marcus Fontoura, *Yahoo! Inc.**

Eugene Shekita, *IBM Almaden Research Center*

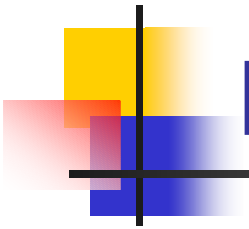
Presented by Pavel Dmitriev

* This work was done while these authors were at IBM Almaden Research Center



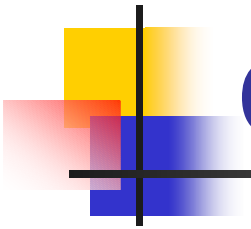
Enterprise Search

- Enterprise Search is search over the data in the company's corporate intranet
  - Intranets can be quite large (millions of pages)
- Recent research shows that employees spend a large percentage of their time searching for information
  - Improvement in quality of enterprise search directly results in increased employee productivity



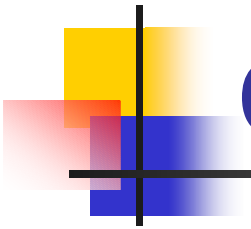
Problems in Enterprise Search

- Users cannot freely create web pages in intranets
  - Algorithms based on link structure analysis do not apply directly
  - The amount of anchor text is limited
  - Anchor text is of “lower quality”
- Lack of good anchor text is one of the major factors affecting search quality



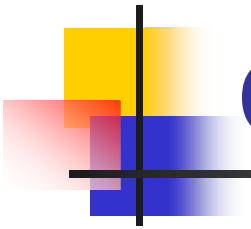
Good News

- No spam in intranets
- Users are willing to cooperate with the search engine



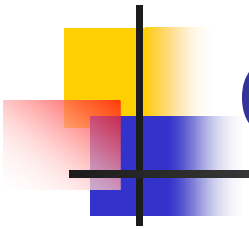
Good News

- No spam in intranets
- Users are willing to cooperate with the search engine
- ➔ *Use user feedback!*



Our Solution

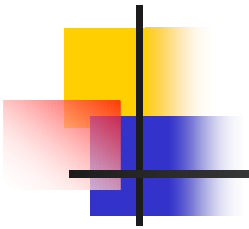
- Use user feedback to make up for the lack of anchor text in intranets
- Focus on a specific form of feedback: “user annotations”



Outline

- Introduction to Annotations
- Collecting Annotations
  - Explicitly
  - Implicitly
- Integrating Annotations into Search Engine
- Experimental Results
- Conclusion & Future Work

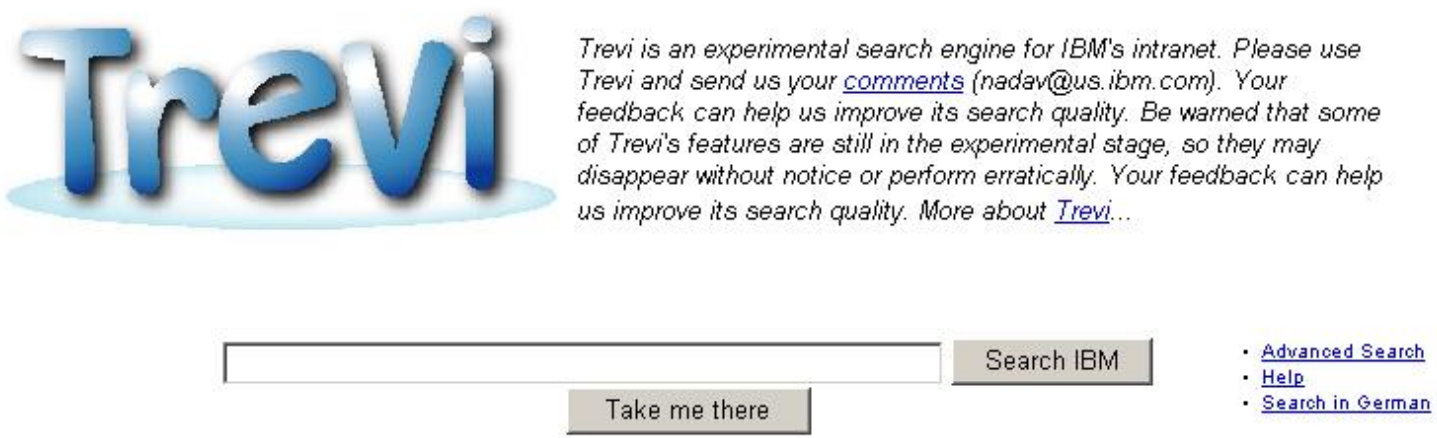
Introduction to Annotations



- An *annotation* is a short description of the contents of a web page
- Intranet users cannot create anchor text → let them create annotations, and use annotations as substitutes for anchor text
- Explicit Annotations: let users enter short descriptions of the pages they see
- Implicit Annotations: derive annotations automatically from users' behavior

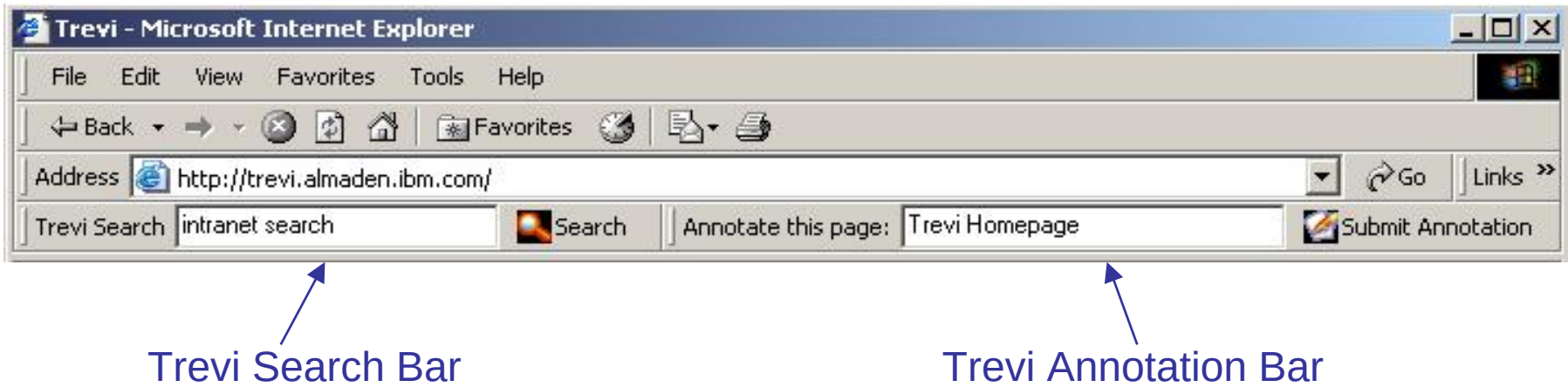
Environment

- Implemented our ideas in the context of IBM Intranet (5.8 million pages, 1.8 million visits per day on average)
- Used Trevi – an experimental search engine for IBM Intranet



Explicit Annotations: Trevi Toolbal

- Users can enter annotations using Trevi Toolbar:

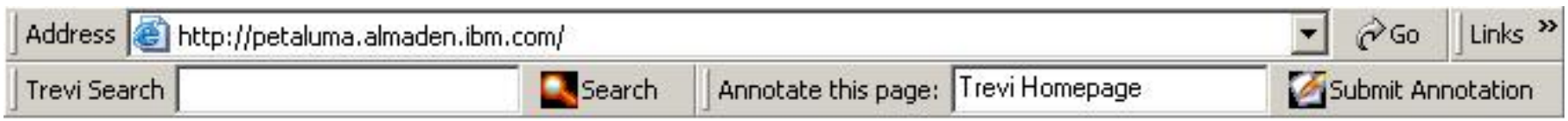


- Trevi Search Bar lets users enter queries to Trevi

- Trevi Annotation Bar lets users enter annotations for the pages they visit

Benefits to the User

- Users directly benefit in two ways
  - They see their annotations in the toolbar while browsing:



  - They see their annotations in the search results:

Trevi

Your annotation: Trevi Homepage

TREVI Trevi is an experimental **search** engine for IBM 's **intranet**. ...Advanced **Search**. Help **Search** in German. TREVI **search Searching** 7,404,416 IBM web pages. ...10/04/03 Rate your **search** results to further enhance Trevi 's **search** quality. ...

petaluma.almaden.ibm.com/

Implicit Annotations

- Idea: use the queries users submit to the search engine as annotations
  - Extract which search results users click on from Trevi log, and use this information to determine which pages are relevant to the query
  - Attach annotations to these pages

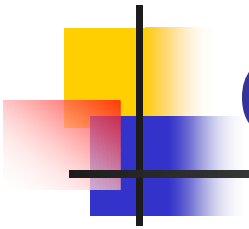
Implicit Annotations

- Idea: use the queries users submit to the search engine as annotations
  - Extract which search results users click on from Trevi log, and use this information to determine which pages are relevant to the query
  - Attach annotations to these pages
- Question: which pages to consider relevant?
  - Strategy 1: Every clicked page is relevant

Implicit Annotations (cont.)

- Session – time-ordered sequence of clicks on search results the user makes for a given query
  - Strategy 2: Attach annotation only to the last clicked page in the session

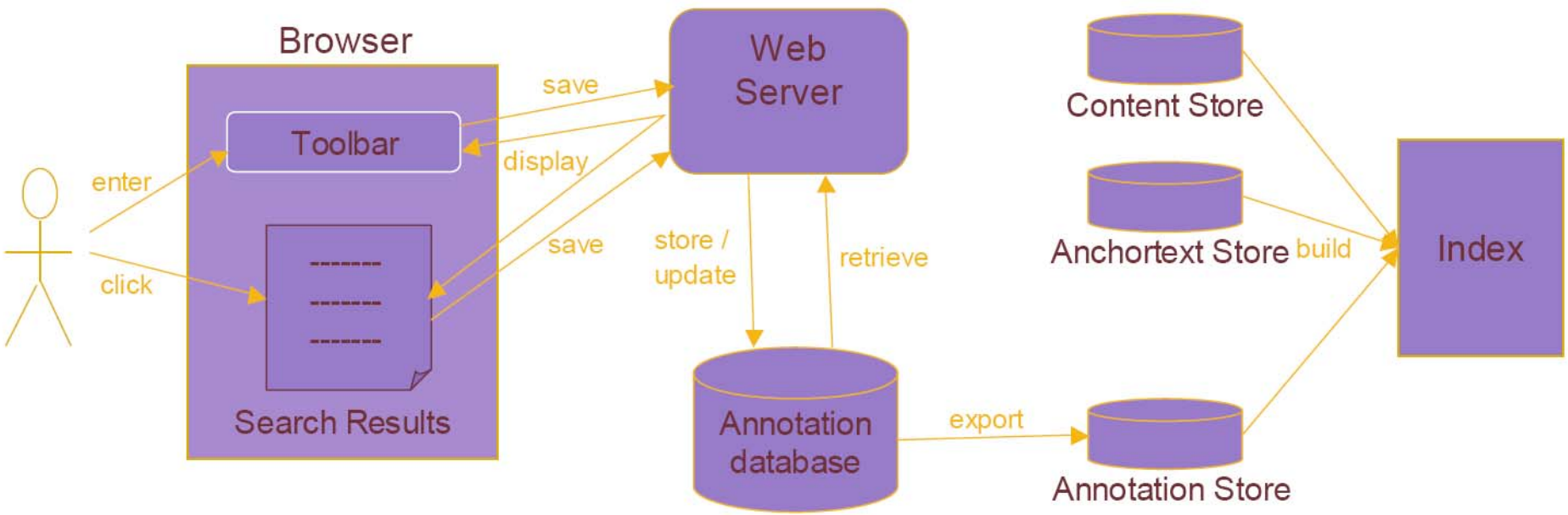
- Query chain – time ordered sequence of queries, executed by the user over a short period of time
  - Annotation = concatenation of all queries from the query chain
  - Strategy 3: Attach annotation to every clicked page
  - Strategy 4: Attach annotation only to the last click in the last session in the query chain



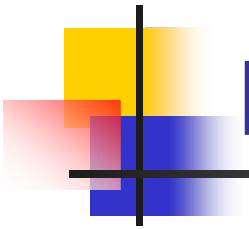
Outline

- Introduction to Annotations
- Collecting Annotations
  - Explicitly
  - Implicitly
- **Integrating Annotations into Search Engine**
- Experimental Results
- Conclusion & Future Work

Integrating Annotations into Trevi



Flow of annotations through the system



Experiments

- Conducted experiments on IBM intranet using Trevi
- Explicit annotations: 158 annotations for 67 pages, 14 in the index
- Implicit annotations: ~12500 for the 1st and 3rd strategies, ~8500 for the 2nd strategy, ~4000 for the 4th strategy

Types of Annotations

- Three common types:
  - Concise descriptions
  - Abbreviations
  - Opinions

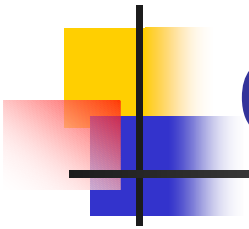
Annotation	Annotated Page
change IBM passwords	Page about changing various passwords in IBM intranet
stockholder account access	Login page for IBM stock holders
download page for Cloudscape and Derby	Page with a link to Derby download
ESPP home	Details on Employee Stock Purchase Plan
EAMT home	Enterprise Asset Management homepage
PMR site	Problem Management Record homepage
coolest page ever	Homepage of an IBM employee
most hard-working intern	an intern's personal information page
good mentor	an employee's personal information page

Impact on Search Quality

- Generated test set from explicit annotations
  - Queries = annotations
  - Correct answers = annotated pages

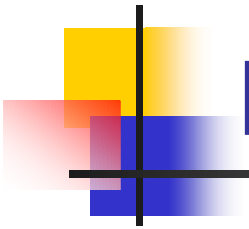
- Measured “Precision at 10”, using search results without annotations as a baseline

Baseline	EA	IA 1	IA 2	IA 3	IA 4
8.9%	13.9%	8.9%	8.9%	9.5%	9.5%



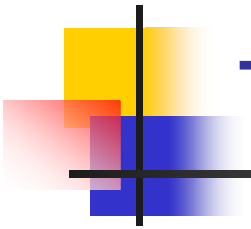
Conclusion

- Described an architecture for collecting annotation and adding them to search indexes
- Presented algorithms for generating implicit annotations from query logs
- Showed preliminary experimental results, suggesting that annotations have potential to improve search quality



Future Work

- More accurate extraction of query chains
- Ranking with annotations
- Detecting irrelevant annotations
- Managing annotations over time
- Other ways of collecting user feedback



Thank you!