Experiences with Selecting Search Engines Using Metasearch

Daniel Dreilinger
MIT Media Laboratory

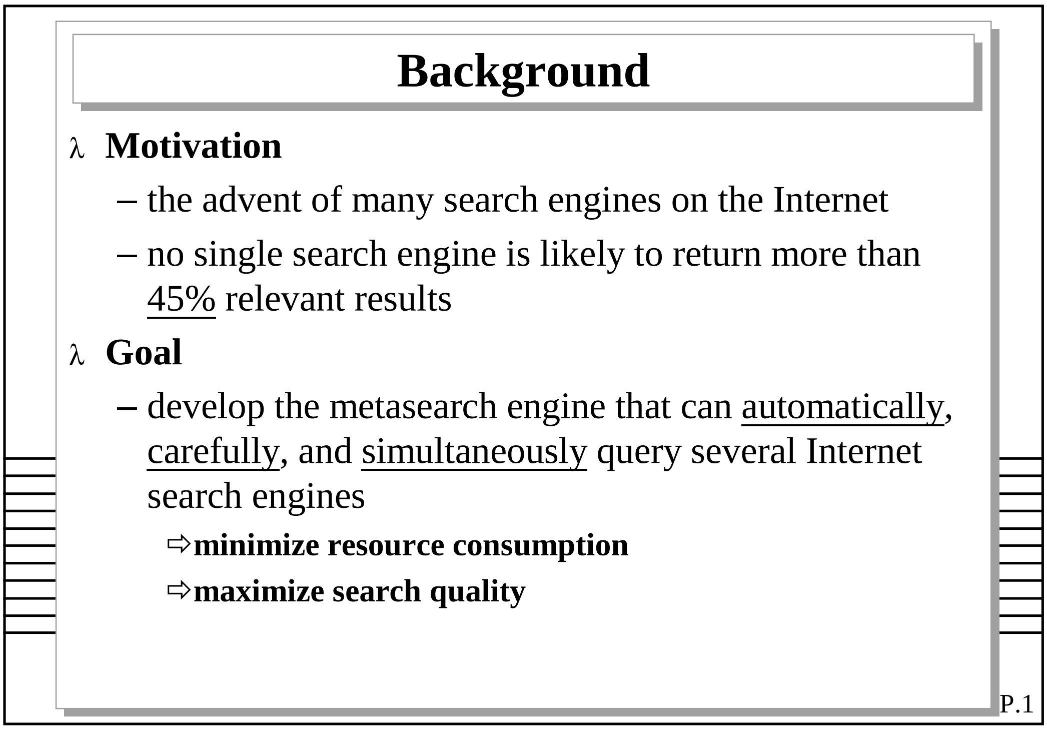
Adele E. Howe
Colorado State University

ACM Transactions on Information Systems, 15(3), 1997
Received 1996.8, Revised 1996.12, Accepted 1996.12

1997.10.23

Outline

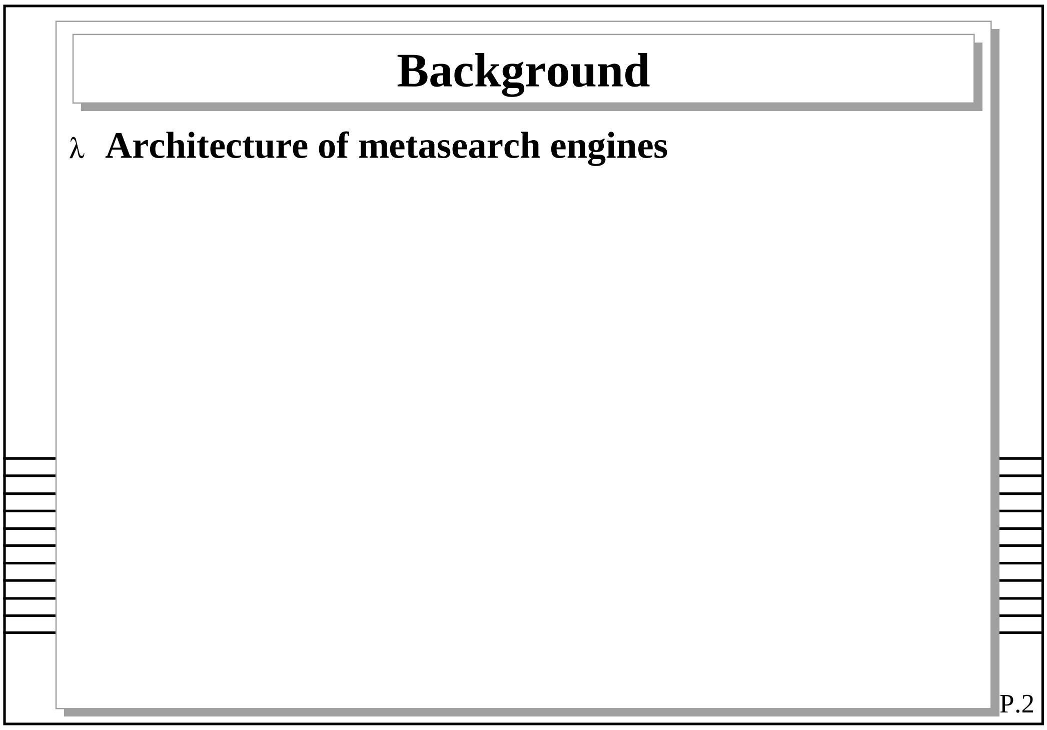
- λ **Background**
- λ **Main Idea**
- λ **Evaluation**
- λ **Conclusion**



Background

- **Motivation**
  - the advent of many search engines on the Internet
  - no single search engine is likely to return more than 45% relevant results
- **Goal**
  - develop the metasearch engine that can automatically, carefully, and simultaneously query several Internet search engines
    - ⇨**minimize resource consumption**
    - ⇨**maximize search quality**

P.1



Background

- **Architecture of metasearch engines**

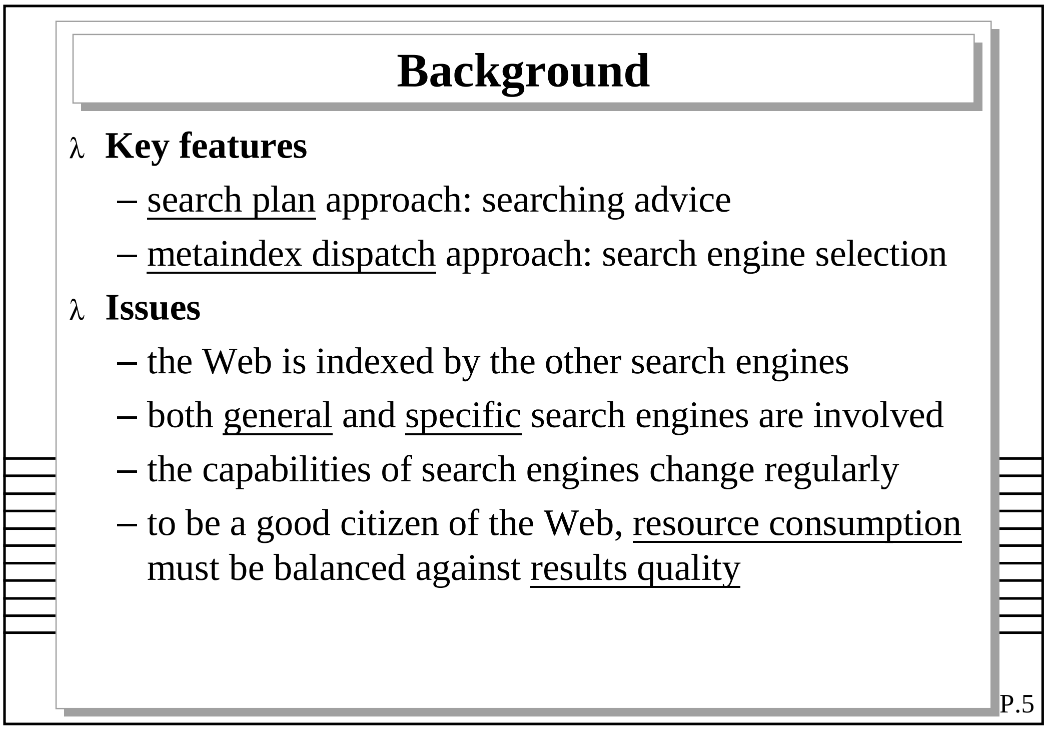
P.2

Background

λ **SavvySearch query form**

Background

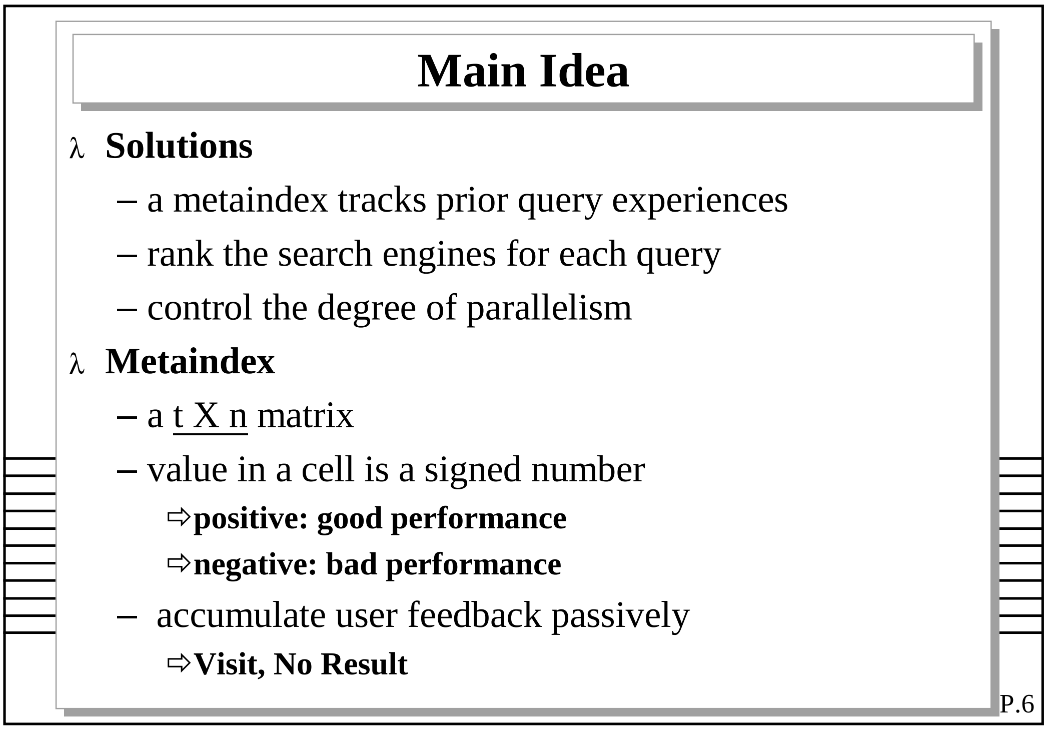
λ **Search plan**



Background

- λ **Key features**
  - – search plan approach: searching advice
  - – metaindex dispatch approach: search engine selection
- λ **Issues**
  - – the Web is indexed by the other search engines
  - – both general and specific search engines are involved
  - – the capabilities of search engines change regularly
  - – to be a good citizen of the Web, resource consumption must be balanced against results quality

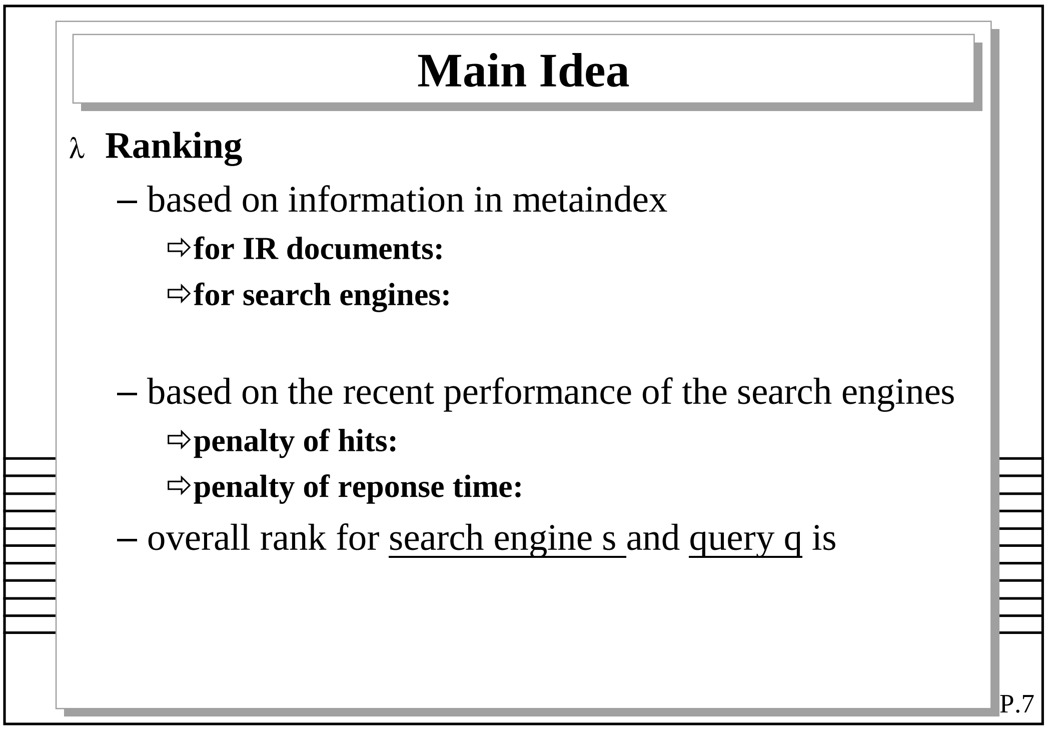
P.5



Main Idea

- λ **Solutions**
  - – a metaindex tracks prior query experiences
  - – rank the search engines for each query
  - – control the degree of parallelism
- λ **Metaindex**
  - – a t X n matrix
  - – value in a cell is a signed number
    - ⇨**positive: good performance**
    - ⇨**negative: bad performance**
  - – accumulate user feedback passively
    - ⇨**Visit, No Result**

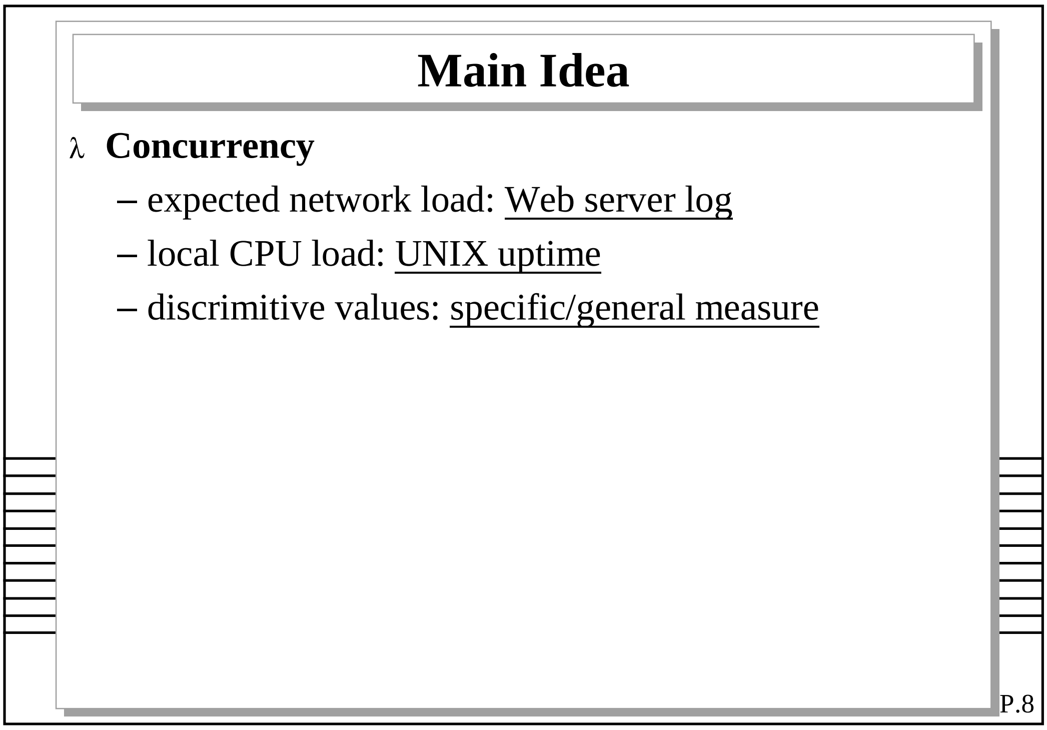
P.6



Main Idea

- λ **Ranking**
  - – based on information in metaindex
    - ⇨**for IR documents:**
    - ⇨**for search engines:**
  - – based on the recent performance of the search engines
    - ⇨**penalty of hits:**
    - ⇨**penalty of reponse time:**
  - – overall rank for search engine s and query q is

P.7



Main Idea

- λ **Concurrency**
  - – expected network load: Web server log
  - – local CPU load: UNIX uptime
  - – discrimitive values: specific/general measure

P.8

Evaluation

- **Experiment I.**
  - Approach A: group order, selection order
    - ⇨**Visit: 2, Self-report: 72%**
  - Approach B: random group order, selection order
    - ⇨**Visit: 1.76, Self-report: 60%**
  - Approach C: group order, random selection order
    - ⇨**Visit: 1.89, Self-report: 65%**
  - Approach D: both random
    - ⇨**Visit: 1.55, Self-report: 60%**
  - quality is significantly improved by the search engine ranking: A>C, B>D

Evaluation

Evaluation

- **Experiment II.**
  - Were the additional knowledge meant better performance?
    - ⇨**No Result: short-term**
    - ⇨**Visit: long-term**
  - If some of the search engines are truly comprehensive, metasearch might be unnecessary

Evaluation

Evaluation

- **Experiment III.**
  - comparison with preprogrammed design
    - ⇨**Visit: 46%, No Result: 12%**
    - ⇨**Visit: 40%, No Result: 14%**
      - ➔Visit: 100~200, No Result: 4~5
  - require considerable experience with a word to surpass programmed approach on the Visit measure, but only a few on the No Result measure

P.13

Conclusion

- **Critiques**
  - general framework of the metasearch engines is not consistent with text
  - terms in mutiword queries may have different contributions with a Visit event
  - the penalty of reponse time may be related to the queries expect for the search engines
- **Difficulties**
  - remote search engines vary in numerous ways
  - user population vary, too

P.14