Click Log Based Evaluation of Link Discovery

David Alexander
Andrew Trotman
Alistair Knott

University of Otago, Dunedin, N.Z.
Overview

• Hypertext as information retrieval
• Link discovery
  • Techniques
  • Automatic and manual assessment
  • Click-based assessment
• Experiments and results
• Conclusions
Hypertext

• A form of information retrieval:
  • Not the same as a search engine.
  • Helps users find what they need even when they don’t know exactly what that is (i.e. browsing).

• But which links will lead to the most productive browsing?
Link Discovery

• Automatic generation of *relevant* hyperlinks

• Various approaches:
  • Link probability (Itakura & Clarke, 2008; Jenkinson et al., 2009)
  • Title matching (Geva, 2008)
  • ...and more “sophisticated” methods. (e.g. Milne & Witten, 2008)

• All assume that link quality is not user-dependent.
Assessment

• Automatic assessment
  • *How well can the algorithm predict the existing links?*

• Manual assessment *(Huang et al., 2009)*
  • *Does the algorithm produce links that meet certain assessable criteria?*

• What exactly are we assessing?
  • The *topical relevance* — not a user-dependent property — of links.
Our Approach

• Get the users to tell us what they want to click on!
  • But not directly...
  • ...through click logs.
• Metrics: Click Count, Reading Time, ...
• Biases: Page Popularity, Anchor Position, ...
Our Data

• Click log:
  • 2,595,572 Wikipedia requests
  • 17,635 unique users

• The INEX Wikipedia Corpus:
  • Snapshot taken in 2005
  • 659,388 articles (451,979 occur in click log)
  • 24,094,179 links (473,510 clicked in log)
Experiment 1: Compatibility

• How are our metric similar to INEX assessment?
  • See what effect INEX-relevance has on our scores using a t-test:
  • The effect is significant (<1%).
  • But these scores are per-link, not per-run.
Experiment 1: Compatibility
Experiment 1: Compatibility
Experiment 1: Compatibility
Experiment 2: Correlation

• How are our metric different from INEX assessment?
  • See which runs do well under each assessment method (using MAP and nDCG):
  • Some do well under INEX, some do well under our metrics, but few do well under both.
  • Correlation is low between manual and click-based assessments; higher, but still not perfect, between automatic and click-based assessments.
Conclusion

- Click-based metrics are a novel way of measuring hypertext.
  - They are not predicted well by existing metrics.
- Standard IR scores (MAP and nDCG) can be calculated for click-based metrics.
- Click-based metrics harness large quantities of unbiased data.
- Further work could see this data used in mainstream (non-research) applications.
Acknowledgements

- My presentation of this paper was made possible with the help of an ADCS travel grant from Google.
Questions?