Improving Website Hyperlink Structure Using Server Logs

Ashwin Paranjape ashwinp@cs.stanford.edu
Robert West west@cs.stanford.edu
Leila Zia leila@wikimedia.org
Jure Leskovec jure@cs.stanford.edu

INTRODUCTION

- Hyperlinks on websites are important for both humans and machines
- But finding and adding missing links is hard
- 66% of links added by editors in a month on Wikipedia are not clicked in the next month
- Can we recommend missing links?
- Links that will be clicked will be useful
- Use server logs as sensor for usefulness

TASK

Web server logs tell us how people use existing links. We can tie these clicks together to infer navigation paths of users.

We consider all unlinked pairs of pages (S,T) and predict how often a link would be clicked if it were added to the site .

Number of times link (S,T) would be clicked = Number of times S is visited × Probability of clicking on a link to T from S

$\#(S,T) = \#(S) \times \Pr(T|S)$

$\#(S)$: Known. $\Pr(T|S)$: Needs to be estimated

REFERENCES

D. Milne and I. H. Witten. Learning to link with Wikipedia. In CIKM, 2008.

T. Noraset, C. Bhagavatula, and D. Downey. Adding high-precision links to Wikipedia. In EMNLP, 2014.

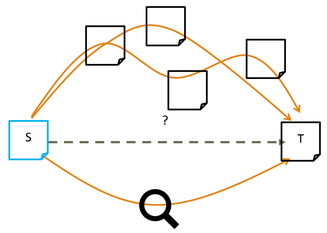
R. West, D. Precup, and J. Pineau. Completing Wikipedia's hyperlink structure through dimensionality reduction. In CIKM, 2009.

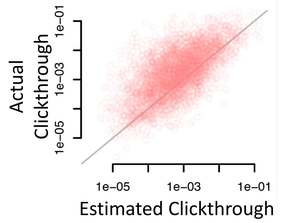
M. Grecu. Navigability in information networks. Master's thesis, ETH Zürich, 2014.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. JASIST, 58(7):1019–1031, 2007.

ESTIMATING CLICKTHROUGH

Hypothesis: Those who take indirect paths from S to T would use the link had it existed





$$\Pr(T|S) \approx \frac{\#(S \to \ldots \to T)}{\#(S)}$$

If only pairwise transitions are available, simulating human behavior using clickthrough probability over existing links gives accurate clickthrough

MODELING LINK INTERACTION

Links are not independent of each other. Adding more links may not preserve individual clickthrough.

Fact: 93% of times, 1 link was taken

Model: User clicks exactly 1 link on a page

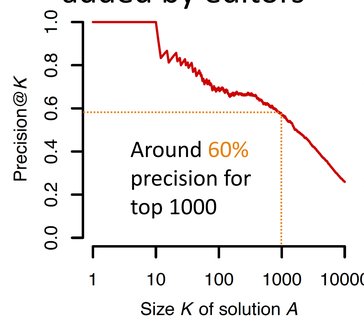
For an S, $\sum_{T:(S,T)\in New} \#(S,T) = \#(S) \times \frac{\sum_{T:(S,T)\in New} \Pr(T|S)}{\sum_{T:(S,T)\in All} \Pr(T|S)}$

Probability that the 1 link taken from a page S is a new link

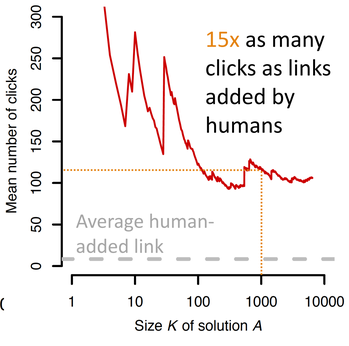
Monotonic for each S and a set of new links
i.e. $\Pr(T|S) \leq \Pr(U|S) \Rightarrow \#(S,T) \leq \#(U,T)$
Greedy for each S: Add links to solution set of S in decreasing order of $\#(S,T)$
Diminishing returns: Later a link is added to the solution set, lesser is its contribution
Marginal gains: Additional #clicks using new links from S after adding (S,T)
Optimization: Greedy marginal gain selection gives optimal solution

EVALUATION: WIKIPEDIA

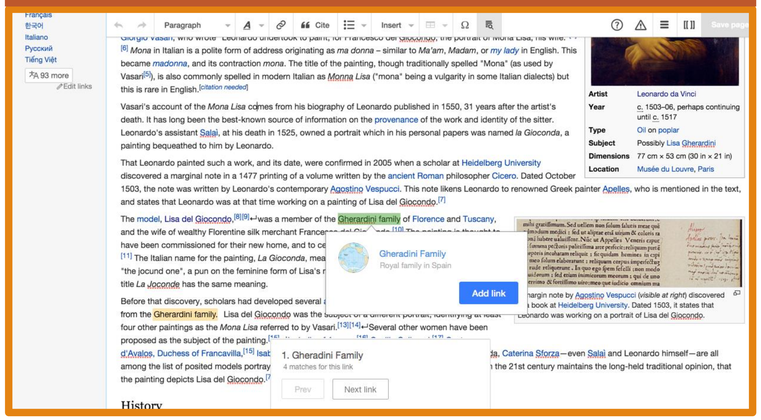
Q1: How many of our top K links were independently added by editors



Q2: How often were our top K links clicked?



IMPACT



We have built an **experimental gadget in Visual editor** in Wikipedia to recommend missing links to an editor who can add it with a single click.

Our method is independent of language, captures recent external events and can indicate missing content. It also worked well on smaller websites.