Mining Missing Hyperlinks from Human Navigation Traces

Bob West *Stanford*
Ashwin Paranjape *Stanford*
Jure Leskovec *Stanford*
Leila Zia *Wikimedia*

Wikimedia Research Showcase
March 25, 2015

Before we get started…

Workshop on Wikipedia, a Social Pedia:
Research Challenges and Opportunities

(Co-located with ICWSM, Oxford, UK, May 26)

http://snap.stanford.edu/wiki-icwsm15

Submit extended abstracts by **Tue, March 31**

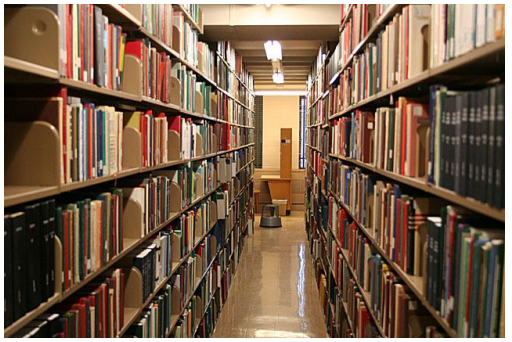
Motivation: The World Wide Web

1. User's perspective

Ingolstadt (German pronunciation: [ˈɪŋgɔlˌʃtat]; Austro-Bavarian [ˈɪŋˌʃtɔːd]) is
Germany. It is located along the banks of the River Danube, in the cen
citizens. It is part of the Munich Metropolitan Area, which has a total pc

The *Illuminati*, a Bavarian secret society, was founded in Ingolstadt in t

vs.



2. Search engine's perspective: PageRank etc.

THE $25,000,000,000* EIGENVECTOR
THE LINEAR ALGEBRA BEHIND GOOGLE

KURT BRYAN† AND TANYA LEISE‡

Abstract. Google's success derives in large part from its PageRank algorithm, which ranks the importance of webpages according to an eigenvector of a weighted link matrix. Analysis of the PageRank formula provides a wonderful applied topic for a linear algebra course. Instructors may assign this article as a project to more advanced students, or spend one or two lectures presenting the material with assigned homework from the exercises. This material also complements the discussion of Markov chains in matrix algebra. Maple and Mathematica files supporting this material can be found at www.rose-hulman.edu/∼bryan.

3. Content provider's perspective

Scenario: Writing a new wiki article

Source prediction

s_1
s_2
s_3
s_4
...
s_{2745}

The ***brdlbrmpft*** is a rare species of pizzazz-based shbuzzle, recently discovered in the Amazon rainforest by brmpftologist Walter Röblimann.

Target prediction

t_1
t_2
t_3
t_4

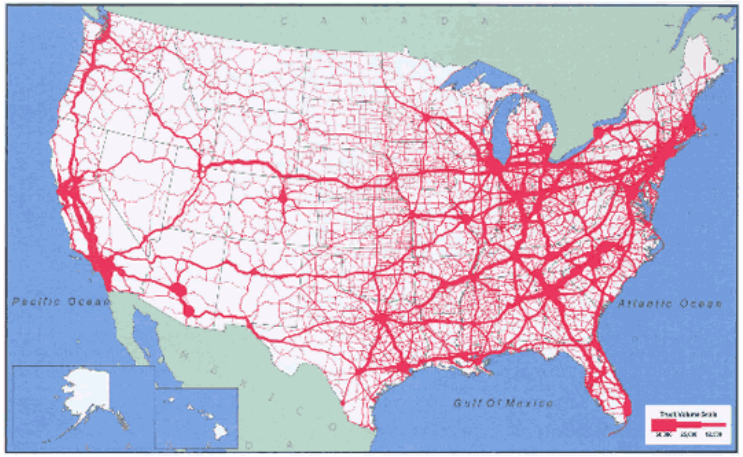
Typical flavor of prior work

1. Build statistical model of existing links
2. Generalize to predict novel links

Static view

"Build new roads that are in tune with the static structure of existing roads."

Pragmatic view



Need roads where the traffic is!

Need links where the traffic is!

Part I

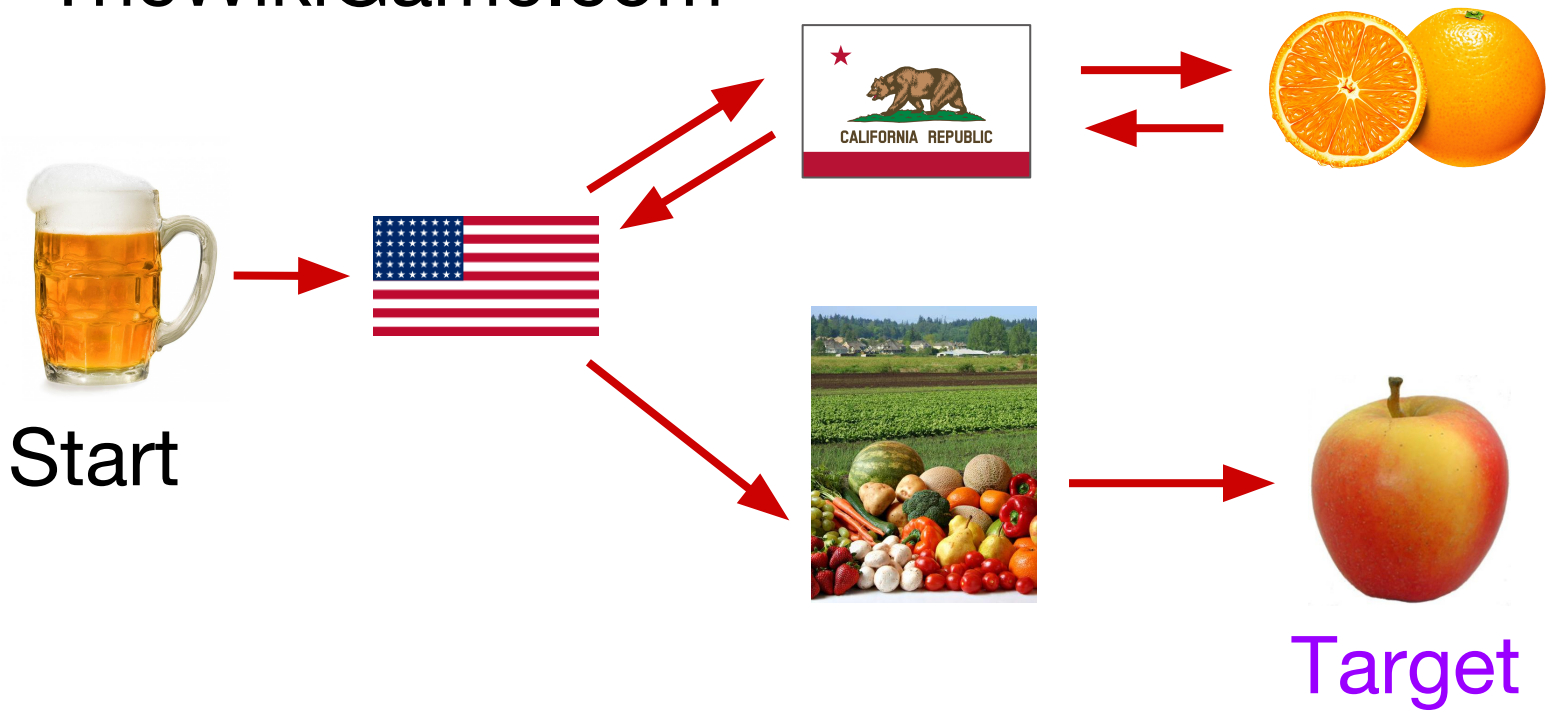
Data collection via human computation

R. West, A. Paranjape, J. Leskovec: Mining Missing Hyperlinks from Human Navigation Traces: A Case Study of Wikipedia, *WWW 2015.*

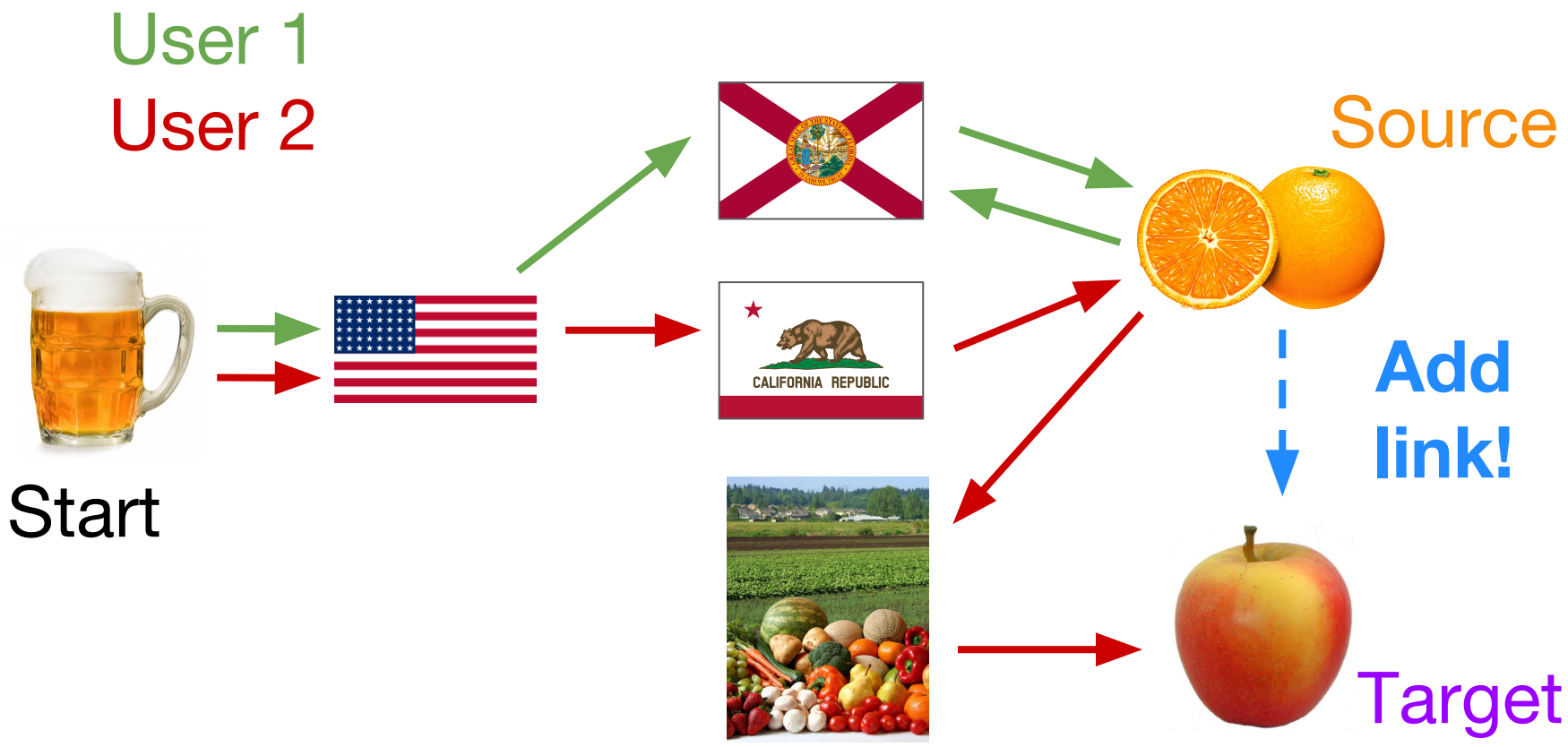
Data set of navigation traces

Collected via human-computation games:

- Wikispeedia.net
- TheWikiGame.com

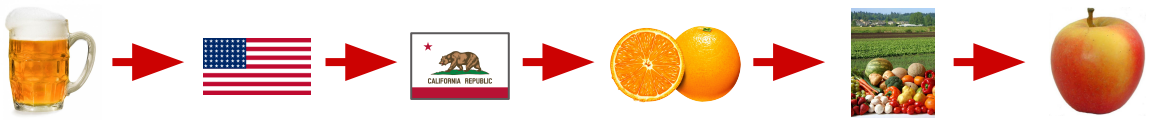


Idea for link-suggestion method



“Algorithm”

1. Given: target t
2. Collect many paths $p = \langle p_0, p_1, \ldots, p_{n-1}, p_n = t \rangle$
3. Initial link candidates: $(p_1, t), \ldots, (p_{n-1}, t)$
4. Filter candidates: keep (s, t) iff
   a. s doesn’t already link to t in current Wikipedia, and
   b. s mentions t in its text
5. Rank remaining candidates, e.g.,
   a. by frequency Pr(user visits s | target is t), or
   b. semantic relatedness between s and t



Evaluation

- See WWW'15 paper* for detailed evaluation
- Gist:
  - Strong baseline:
    - Check *all* articles *s* for mentions of target *t* and rank by semantic relatedness of (*s*, *t*)
  - Our version 1 (better than baseline):
    - Check only small set of candidates from paths to *t* for mentions of *t* and rank by sem. relatedness
  - Our version 2 (as good as baseline):
    - Check only small set of candidates from paths to *t* for mentions of *t* and rank by frequency Pr(*s* | *t*)

* Robert West, Ashwin Paranjape, and Jure Leskovec: Mining Missing Hyperlinks from Human Navigation Traces: A Case Study of Wikipedia, *WWW 2015.*

Discussion

Advantages:

+ Addressing source prediction
+ Optimizing right objective (navigability)
+ No model of link structure needed
+ Simple filtering + ranking are enough
+ Because heavy lifting is done by humans

Disadvantage:

– Need to know navigation target

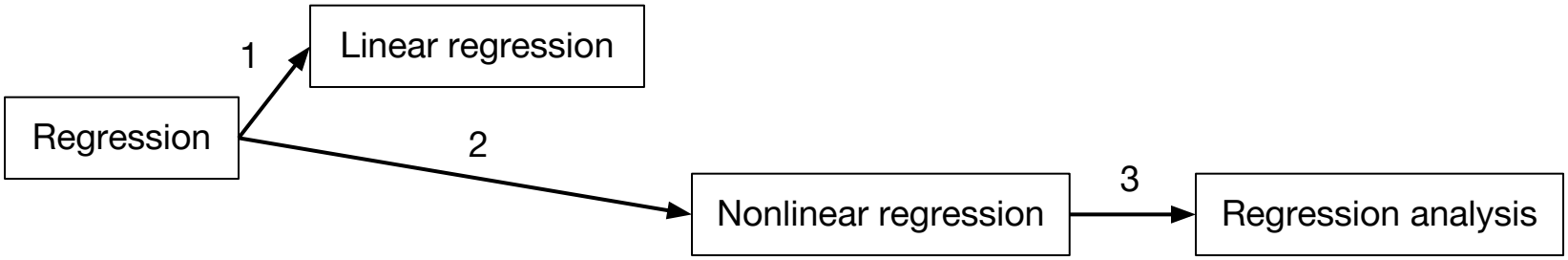
Part II

Data collection via Wikimedia server logs

Ongoing work. Collaborators:
Ashwin Paranjape, Jure Leskovec, Leila Zia

Ongoing work: Adapting our method “for the wild”

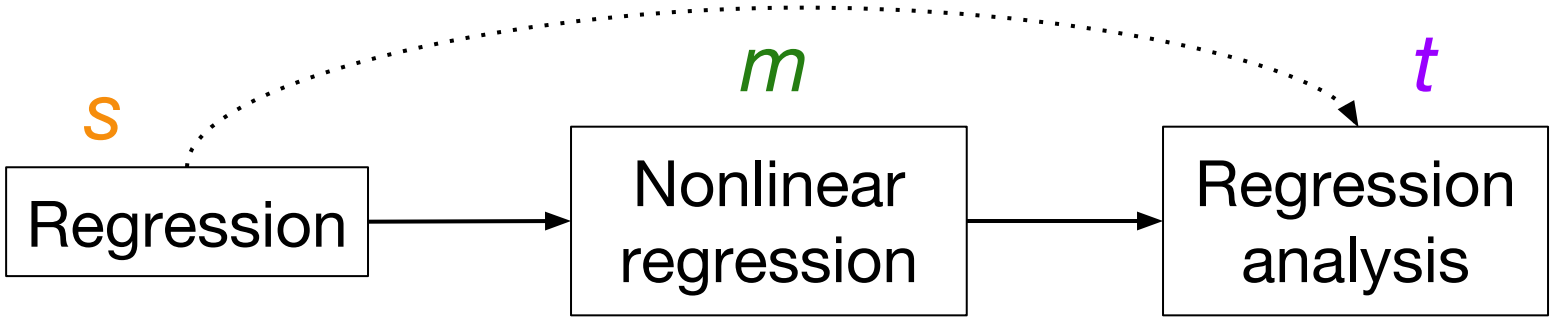
- Raw data: Wikipedia webserver logs
- String pageviews together into navigation traces, e.g.



- Extracted navigation traces (trees) for 50 largest language versions
- Number of English trees for 1 month: 2.4 B

Algorithm

- Challenge: Navigation target unknown (or even non-existent)
- Idea: Close triangles if justified by data



Suppose
Pr(*t* | *m*) = 5%
Pr(*t* | *s*, *m*) = 34%

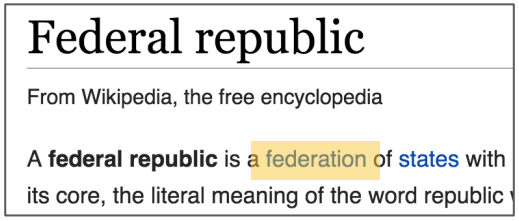
→ *s* drives users to *t* → Suggest link (*s*, *t*)

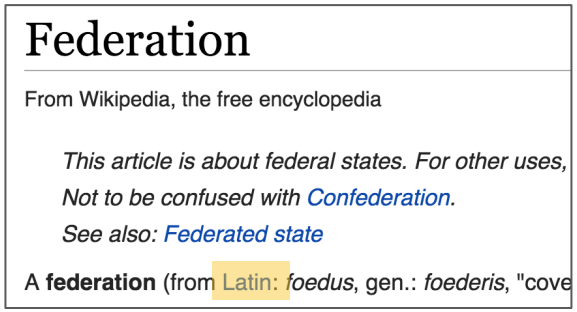
“Results”

s	*m*	*t*	Pr(*t* | *m*)	Pr(*t* | *s*, *m*)	Diff.	Comment
2014–15 Bundesliga	FC Bayern Munich	2014–15 FC B.M. season	5% (4,599 / 91,948)	23% (451 / 1,946)	18%	Good suggestion
Wojciech Szczęsny	Poland national football team	Artur Boruc	2% (571 / 12,803)	20% (122 / 536)	18%	Both are Polish goalkeepers, but *s* doesn’t mention *t*
2008 World Series	2007 World Series	2006 World Series	9% (241 / 2,694)	80% (200 / 249)	71%	Iterating through list*
Federal republic	Federation	Latin	2% (389 / 23,702)	20% (368 / 1,848)	18%	Following first links**

*

**





Discussion

- Promising results even without (known) target
- Applications beyond link suggestion
  - Mining typical browsing patterns
    - Following first links
    - Iterating through lists
    - …
  - Automatic page restructuring
    - If *s* doesn't mention *t*, suggest it do
    - Important links should appear early
  - Automatically constructing reading lists and curricula
- Bot in preparation

Summary

- Motivation: Links are key, but are often missing
- Here: Simple method to find missing links
- How? Mine usage logs to guess which non-existent links were expected to exist by users
- Version 1: Human computation (done)
- Version 2: Wikipedia server logs (ongoing)

Thanks!

west@cs.stanford.edu