The Chinese University of Hong Kong

PAGESIM: A NOVEL LINK-BASED MEASURE OF WEB PAGE SIMILARITY

Zhenjiang Lin, Michael R. Lyu, and Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR
{zjlin, lyu, king}@cse.cuhk.edu.hk

1. Motivation

Finding similar/related pages on the Web!

1. Finding similar Web pages
2. Web document classification
3. Other applications

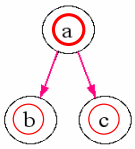
2. What's PageSim

- A Novel link-based Web page similarity measure.
- Can measure similarity between any two Web pages (some existing measures, such as SimRank, cannot always do this).
- The inherent parallelism property of PageSim makes it suitable for distributed computing environment.

3. Intuitions

Intuitions Behind PageSim

- Web page *a* linking to Web page *b* implies that *a* propagates part of its similarity (feature) information to *b* through hyperlinks.
  - On the right figure, Web page *a* propagates its similarity information to its neighbors *b* and *c*.
  - Web pages *b* and *c* carry similarity information of Web page *a*, so all of them are similar Web pages.



- Each Web page caries its own similarity information.
- Authoritative / importance Web pages carry higher similarity information than un-authoritative / unimportant Web Pages.
- The more common similarity information two Web pages have, the more similar they are.

4. PageSim

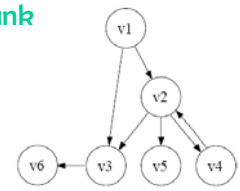
PageSim: PageRank score propagation

- PageRank score is adopted to represent the value of similarity information of a Web page.
- Each Web page propagates its own similarity information to its outlink neighbors, while receiving and propagating similarity information from inlink neighbors at the same time.
- After the propagation, the PageSim score of any two pages is the sum of their common similarity information.
- PageSim inherits parallelism property, since each Web page propagates similarity information independently.

5. Illustration

A simple example: PageSim vs SimRank

Note: top(*v*, *t*) denotes the top *t* similar Web pages to page *v* (excluding *v*).



The results produced by PageSim is

$$top(v_1,2)=\{v_3,v_6\}, \quad top(v_2,2)=\{v_4,v_{3,6}\},$$
$$top(v_3,2)=\{v_6,v_2\}, \quad top(v_4,2)=\{v_2,v_{3,5,6}\},$$
$$top(v_5,2)=\{v_{2,3,4,6},v_1\}, \quad top(v_6,2)=\{v_3,v_2\}.$$

The results produced by SimRank is

$$top(v_1,2)=\{\}, \quad top(v_2,2)=\{v_3,v_6\},$$
$$top(v_3,2)=\{v_{4,5},v_2\}, \quad top(v_4,2)=\{v_5,v_3\},$$
$$top(v_5,2)=\{v_4,v_3\}, \quad top(v_6,2)=\{v_{2,4,5},v_3\}.$$

6. Discussion

The results of PageSim and SimRank are different.

- SimRank: no page similar to v1.
- PageSim: v3 is most similar to v1,

  In fact, v1 linking to v3 does imply that v1 and v3 are related.
- SimRank: v4 is not similar to v2.
- PageSim: v4 is similar to v2.

  In fact, v2 and v4 are similar, because they link to each other.
- SimRank: v3 is most similar to v2 for their having a common inlink page v1.
- PageSim: v4 is most similar to v2 for their linking to each other.

We believe PageSim is more reasonable in this situation because the "linking to each other" relationship does imply stronger similarity than that of "common inlink" relationship.