

Improved Algorithms for Topic Distillation...

by Krishna Bharat and Monika Henzinger

David Deschenes
November 5, 2002
WWW Search Engines

Topic Distillation vs. Topic Exploration

» Topic Distillation

- Assumes a set of pages dealing with a specific topic
- Finds pages within that set which are of high quality

» Topic Exploration

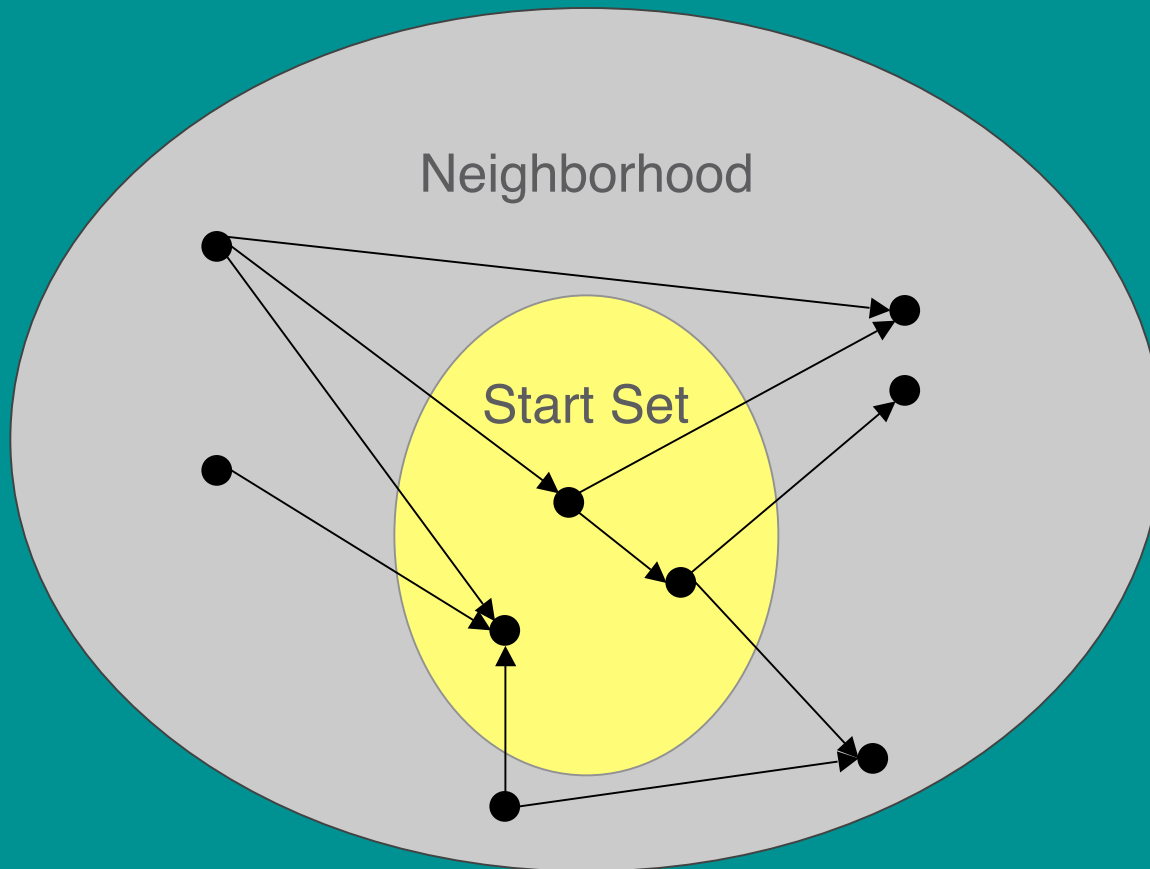
- Locate a set of pages dealing with a specific topic
- May be a preliminary step to topic distillation

Review of Kleinberg's Algorithm

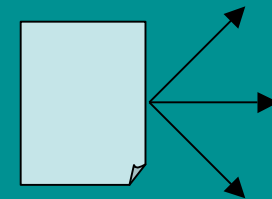
» Connectivity Analysis

- An example of topic distillation
- Builds a neighborhood of topic specific pages
- Hubs point to pages with relevant content
- Authorities contain relevant content
- Iteratively compute hub and authority scores

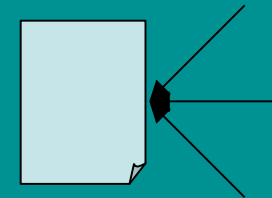
Topic Specific Page Set



Hub



Authority



Kleinberg's Algorithm

- (1) Let N be the set of nodes in the neighborhood
- (2) For all $n \in N$ let $H(n)$ be its hub rank and let $A(n)$ be its authority rank
- (3) Initialize $H(n)$ and $A(n)$ to 1 for all $n \in N$
- (4) While the vectors H and A have not converged
- (5) For all $n \in N$, $A(n) = \sum_{(n',n) \in N} H(n')$
- (6) For all $n \in N$, $H(n) = \sum_{(n',n) \in N} A(n')$
- (7) Normalize the H and A Vectors

Problems with Kleinberg's Algorithm

- » Mutually Reinforcing Relationships
 - Give undue weight to the opinion of one individual
- » Automatically Generated Links
 - Violates assumption that links represent opinions
- » Non-relevant Documents
 - Documents can be highly-linked yet off-topic

Solutions

- » Introduce content analysis to the process
 - Compute page relevance
- » Improve the connectivity analysis
 - Handle mutually reinforcing relationships
 - Weight rankings using page relevance
- » Improve the start set of pages
 - Remove non-relevant pages
 - Reduce the effect of automatically generated links

Introducing Content Analysis

- » Compute relevance measures for pages
 - Make the query more representative of the topic
 - Relevance is the similarity of the page to the topic
- » Uses of relevance measures
 - Modify connectivity analysis algorithm
 - Facilitate improving the start set of pages

Improved Connectivity Analysis

- » Distribute influence across edges from one host
 - If there are k edges from many pages on one host to a single page on another host then each edge is given an authority weight of $1/k$
 - If there are l edges from a single page on one host to many pages on another host then each edge is given a hub weight of $1/l$
- » Weigh influence using page relevance measures

Modified Kleinberg Algorithm

- (1) Let N be the set of nodes in the neighborhood
- (2) For all $n \in N$ let $H(n)$ be its hub rank, let $A(n)$ be its authority rank, and let $W(n)$ be its relevance measure
- (3) Initialize $H(n)$ and $A(n)$ to 1 for all $n \in N$
- (4) While the vectors H and A have not converged
- (5) For all $n \in N$,
$$A(n) = \sum_{(n',n) \in N} H(n') \times \text{auth_wt}(n',n) \times W(n')$$
- (6) For all $n \in N$
$$H(n) = \sum_{(n',n) \in N} A(n') \times \text{hub_wt}(n,n') \times W(n')$$
- (7) Normalize the H and A Vectors

Improved Start Set of Pages

- » Remove non-relevant documents by
 - Setting a relevance threshold
 - Median Weight
 - Start Set Median Weight
 - Fraction of Maximum Weight
 - Setting a document degree threshold
 - $4 \times (\text{In Degree}) + (\text{Out Degree})$
 - Iteratively removing non-relevant highly-ranked pages

Testing Procedure

- » The techniques just discussed were combined in various ways to produce 8 unique ranking algorithms
- » Each algorithm was performed on the same set of 28 topics

Evaluation Procedure

- » Page relevance determined by a panel of 3 volunteers
- » Volunteers reviewed pools of the top pages from all of the algorithms for each of the topics
- » Computed relative recall and precision, for both hubs and authorities, at 5 and 10 pages retrieved

Evaluation Results

- » In every case, the best algorithm improves precision by at least 45% when compared to Kleinberg's algorithm
- » Simply adding the ability to handle Mutually Reinforcing Relationships yielded the most benefit
- » Recall statistics don't tell us much as the algorithms were only evaluated out to 10 pages retrieved

Questions

- » Would the evaluation results be more valuable if the evaluation had been performed past the first ten pages retrieved?
- » Was the evaluation procedure justified and fair?