Link Analysis: HITS

CMSC 498J: Social Media Computing
Department of Computer Science
University of Maryland
Spring 2016

Hadi Amiri
hadi@umd.edu
Lecture Topics

• The Problem of Ranking
• HITS
  ▫ The Principle of Repeated Improvements
• Spectral Analysis of HITS
Ranking Problem

- **Web Ranking Problem:**
  - Given the Web and a query, rank Web pages with respect to the query such that the most relevant pages to the query appear higher in the ranked list.
Ranking Problem - Cnt.

- Information Retrieval
  - Started in 1960s!
  - *search* repositories of newspapers, papers, patents, legal abstracts, in response to *queries*
Ranking Problem - Cnt.

- **Web Search**
  - **Dynamics**
    - September 11, 2001!
    - Query:
      - *World Trade Center*
    - Results:
      - descriptive pages about the building
      - Periodically collected & indexed Web pages
  - **Abundance**
    - Large number of relevant documents
    - Users look at a few of them
    - Find the most relevant ones!
Ranking Problem - Cnt.

- **Web Search**
  - Complex vs. Short Queries

- **Synonymy issue**
  - words with the same meaning
    - Scallions and Green onions

- **Polysemy issue**
  - words with multiple meanings
    - Jaguar: automobiles, football players, Apple’s OS.
HITS

• Hyperlink-Induced Topic Search (HITS)
  ▫ A Link analysis algorithm for ranking Web pages.

• Links are essential for ranking
  ▫ In-links could be considered as endorsements!

• In aggregate, if a page receives many links from other (relevant) pages, then its is receiving collective endorsement!

HITS- Cnt.

• How to operationalize such endorsement process?
  ▫ Collect a large sample of pages relevant to the given query (e.g. “newspapers”)
    • Use text-based Information Retrieval
  ▫ Pages in this sample “vote” / “endorse” through their links
    • A page is more important if it receives more votes (endorsement or in-links)
HITS- Cnt.

• Q: newspaper

**Experts** vote for many authoritative pages!
• these pages may have some sense of where the good answers are
• Score them highly
HITS- Cnt.

• Interesting Web pages fall into two categories:
  1. **Authorities** that are pages containing relevant information
     ▫ Newspapers homepages
     ▫ Universities homepages
  2. **Hubs** are pages that link to authorities
     ▫ Lists of newspapers
     ▫ Directories
HITS- Cnt.

• A **good hub**?
  ▫ links to many good authorities

• A **good authority**?
  ▫ is linked from many good hubs

• We use two scores for each node
  ▫ **Hub score** and **Authority score**
HITS- Cnt.

• **Authority Score:**
  - For each page $p$ is the sum of the hub scores of all pages that point to it.

• **Hub Score:**
  - For each page $p$ is the sum of the authority scores of all pages that it points to.
Let’s assume a bipartite graph with hub and authority sets for now.

Note that in practice the graph is not bipartite.
• Q: newspaper
  1. Computing authority scores
     • Hub scores are set to 1 initially!
• Q: newspaper
  2. Computing Hub scores
HITS - Cnt.

Q: newspaper

3. Computing authority scores
HITS- Cnt.

- Q: newspaper
  - Repeat for some iterations!
  - Normalize the scores!
HITS- Cnt.

• Authority Update Rule:
  ▫ For each page \( p \), update \( auth(p) \) to be the sum of the hub scores of all pages that point to it.

• Hub Update Rule:
  ▫ For each page \( p \), update \( hub(p) \) to be the sum of the authority scores of all pages that it points to.
HITS- Cnt.

Algorithm
1. Set all hub scores and authority scores to 1.
2. Choose a number of steps $k$.
3. Perform a sequence of $k$ hub-authority updates:
   1. First apply the Authority Update Rule to the current set of scores.
   2. Then apply the Hub Update Rule to the resulting set of scores.
4. Normalize authority and hub scores
• What happens if we run HITS for larger and larger values of $k$?
  ▫ The normalized values converge to limits as $k$ goes to infinity!
  • Values stabilize; further updates lead to smaller and smaller changes in the values we observe!
Spectral Analysis of HITS

• Let $M_{ij} \in n \times n$
  ▫ denote the adjacency matrix of our Web page sample!
    • If there is a directed edge from page $i$ to page $j$
      • $M_{ij} = 1$
    • Otherwise
      • $M_{ij} = 0$

\[
\begin{bmatrix}
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}
\]
Spectral Analysis of HITS- Cnt.

- Let $M_{ij} \in n \times n$
  - denote the adjacency matrix of our Web page sample!
- Represent Hub and Authority scores by two n-dimension vectors
  - **Hub**: $h \in n \times 1$
    - $h_i$ represents the hub score of node $i$
  - **Authority**: $a \in n \times 1$
    - $a_i$ represents authority score of node $i$
Spectral Analysis of HITS- Cnt.

- **Hub Update Rule:**
  - For each page $i$, update $h_i$ to be the sum of the authority scores of all pages that it points to.

\[
h_i \leftarrow M_{i1}a_1 + M_{i2}a_2 + \cdots + M_{in}a_n,
\]

\[
h \leftarrow Ma.
\]
Spectral Analysis of HITS- Cnt.

• **Authority Update Rule:**
  - For each page $i$, update $a_i$ to be the sum of the hub scores of all pages that point to it.

\[ a_i \leftarrow M_{1i} h_1 + M_{2i} h_2 + \cdots + M_{ni} h_n. \]

\[ a \leftarrow M^T h. \]
Spectral Analysis of HITS- Cnt.

• Unwinding the $k$-step hub-authority computation
  ▫ Let
    • $a^{<o>}$ initial vector of authority scores
    • $h^{<o>}$ initial vector of hub scores
  ▫ Compute $a^{<k>}$ and $h^{<k>}$ vectors!
Spectral Analysis of HITS- Cnt.

\[ a^{(1)} = M^T h^{(0)} \]

\[ k=1 \]
\[ h^{(1)} = M a^{(1)} = M M^T h^{(0)} \]

\[ a^{(2)} = M^T h^{(1)} = M^T M M^T h^{(0)} \]

\[ k=2 \]
\[ h^{(2)} = M a^{(2)} = M M^T M M^T h^{(0)} = (M M^T)^2 h^{(0)} \]

\[ a^{(3)} = M^T h^{(2)} = M^T M M^T M M^T h^{(0)} = (M^T M)^2 M^T h^{(0)} \]

\[ k=3 \]
\[ h^{(3)} = M a^{(3)} = M M^T M M^T M M^T h^{(0)} = (M M^T)^3 h^{(0)} \]
Spectral Analysis of HITS-Cnt.

- Unwinding the $k$-step hub-authority computation

\[
a^{(k)} = (M^T M)^{k-1} M^T h^{(0)}
\]

\[
h^{(k)} = (MM^T)^k h^{(0)}.
\]

- Do they converge to stable values?
Spectral Analysis of HITS- Cnt.

• Magnitude of hub and authority scores grow with each update.

• They only converge when we normalize them!

• In fact, it is the directions of the hub and authority vectors that are converging
  ▫ Why?
Spectral Analysis of HITS- Cnt.

• There are normalization constants $c$ and $d$ so that the following vectors converge to limits as $k$ goes to infinity.

\[
\frac{h^{(k)}}{c^k} \quad \frac{a^{(k)}}{d^k}
\]

• Let’s focus on hub vectors (same for authority vectors)!

\[
h^{(k)} = (MM^T)^k h^{(0)}.
\]

\[
\frac{h^{(k)}}{c^k} = \frac{(MM^T)^k h^{(0)}}{c^k}
\]
Spectral Analysis of HITS- Cnt.

- This

\[
\frac{h^{(k)}}{c^k} = \frac{(MM^T)^k h^{(0)}}{c^k}
\]

- converges to limit \( h^{(*)} \), thus
  - the direction of \( h^{(*)} \) at the limit should not change when multiplied with \( MM^T \)
  - Though it’s length may change by a factor of \( c \).

\[
(MM^T)h^{(*)} = ch^{(*)}.
\]
Spectral Analysis of HITS- Cnt.

- **Definition 1**: vector $v$ is an eigenvector of matrix $X$ if:
  - $Xv = \lambda v$
  - $v$ an eigenvector of $X$ and $\lambda$ is its eigenvalue.

- $h^{<\ast>}$ has to be an eigenvector of $MM^T$.

\[(MM^T)h^{<\ast>} = \chi h^{<\ast>}\].
• **Definition 2:** Any $n \times n$ symmetric matrix has a set of $n$ eigenvectors that are unit vectors and mutually orthogonal
  ▫ they form a basis for the space $\mathbb{R}^n$. 

\[ X = X^T \]
Spectral Analysis of HITS- Cnt.

- $MM^T$ is symmetric!
  - Thus $MM^T$ has $n$ eigenvectors $v_1, v_2, ..., v_n$ with corresponding eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$
    - Let’s assume that: $|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_n|$
  - Given any vector $u$, a good way to think about $(MM^T)u$ is to first write $u$ as a linear combination of $(MM^T)$’s eigenvectors!
    - $h^{(k)} = (MM^T)^k h^{(0)}$
    - $h^{(k)} = (MM^T)^k (q_1 v_1 + ... + q_n v_n) =
      q_1 (MM^T)^k v_1 + ... + q_n (MM^T)^k v_n =
      q_1 (\lambda_1)^k v_1 + ... + q_n (\lambda_n)^k v_n$
Spectral Analysis of HITS - Cnt.

- \( h^{<k>} = q_1 (\lambda_1)^k \mathbf{v}_1 + q_2 (\lambda_2)^k \mathbf{v}_2 + ... + q_n (\lambda_n)^k \mathbf{v}_n \)
  - \( |\lambda_1| > |\lambda_2| > = ... > = |\lambda_n| \)

- \( h^{<k>} / (\lambda_1)^k = q_1 \mathbf{v}_1 + q_2 (\lambda_2/\lambda_1)^k \mathbf{v}_2 + ... + q_n (\lambda_n/\lambda_1)^k \mathbf{v}_n \)
  - What does happen if \( k \) go to infinity?
    - every term except the first goes to 0!
    - Therefore, \( h^{<k>} / (\lambda_1)^k \) converges to \( q_1 \mathbf{v}_1 \)

- Remaining steps:
  - Relaxing \( |\lambda_1| > |\lambda_2| > = ... > = |\lambda_n| \)
  - See book: pages 372-374
Questions?
Reading

• Ch.14 Link Analysis and Web search [NCM]