Path Analysis
References: Ch.10, Data Mining Techniques By M.Berry, and G.Linoff
http://www9.org/w9cdrom/68/68.html

Dr Ahmed Rafea
Outline

• Introduction
• Link Analysis
• Path Analysis Using Markov Chains
• Applications
• With the rapid growth of the WWW, it is almost impractical for individual users to navigate effectively through many of the web documents.
• The most obvious and prominent methods are search engines to access information from the WWW.
• While search tools and directories are very useful, they are seldom efficient for the user to "navigate" through a set of related/connected pages.
• There are alternate approaches that are currently adopted to address the navigation problem.
  – The identification of important hubs and authorities which are important sites that the user might want to browse
  – The agent assisted navigation in which, the system suggests links that the user can follow during the process of browsing.
  – The tour generation wherein the system generates a tour which takes the user from one link to another.
Introduction(2)

• Two approaches will be presented
  – Link Analysis which is based on Graph Theory and is quite effective in Identifying authoritative sources of information on the WWW.
  – Markov chain which is a probabilistic approach to the problem of web link sequence modeling, analysis and prediction.
Link Analysis (1)

- Web pages = nodes
- Hyperlinks = edges
- Spiders & Web crawlers updating
- Hub – a page that links to many authorities
- Authority – a page that is linked to by many hubs
- Authority versus mere popularity
  - Rank by number of unrelated sites linking to a site yields popularity
  - Rank by number of subject-related hubs that point to them yields authority
  - Helps to overcome the situation that often arises in popularity where the real authority (e.g., Home Page) is ranked lower because of lack of popularity of links to it
Link Analysis (2)

• Kleinberg’s Algorithm
  – Creating the root set
    • Using a text-based search engine to find pages containing the search string
  – Identifying the candidates
    • The root set is expanded to include pages that point to or are pointed by pages in the root set
  – Ranking hubs and authorities
    • The candidates are ranked iteratively according to their strength as hubs (have links to many authorities) and authorities (have links from hubs)
Link Analysis (3)
Creating the root set

• Conduct content based search using a text string
• The main idea of search engines is to remove stop words from the query, stem the remaining words and match them against the web pages content.
• There are many variations of matching
• The top n documents are used to establish the root set
• A typical value of n is 200
Identifying the candidates

• Locate pages that the root set pointing to
• Locate subset of pages that are pointing to the root set pages using the URL of the root set pages as the search string.
• The reason for using only subset of pages (d pages) that are pointing to the root set, is to guard against bringing in an unmanageable number of sites.
• A typical value of d is 50
Ranking Hubs and Authorities

- Initialize the A (authority indicator), and H (Hub indicator) for each page by 1
- The A value for each page is updated by adding up the H values of all pages pointing to it.
- The A values for all pages are then normalized so that the sum of their squares equal 1
- The H value for each page is updated by adding up the A values of all pages that this page is pointing to
- The H values for all pages are then normalized in the same way as A normalization
- The process is repeated until A and H values converge
- The pages that end with the highest H values are the strongest Hubs and the ones that end with the highest A values are the strongest authorities.
Markov Chain Models for Link Prediction (1)

- A discrete Markov chain model can be defined by the tuple \( \langle S, A, \lambda \rangle \). \( S \) corresponds to the state space, \( A \) is a matrix representing transition probabilities from one state to another. \( \lambda \) is the initial probability distribution of the states in \( S \). The fundamental property of Markov model is the dependency on the previous state. If the vector \( s[t] \) denotes the probability vector for all the states at time \( 't' \), then:

\[
s(t) = s(t-1) A
\]

- If there are \( 'n' \) states in our Markov chain, then the matrix of transition probabilities \( A \) is of size \( n \times n \).

- Markov chains can be applied to web link sequence modeling. In this formulation, a Markov state can correspond to any of the following:
  - URI/URL
  - HTTP request
  - Action (such as a database update, or sending email)

- Each element of the matrix \( A[s, s'] \) can be estimated as follows:

\[
A(s, s') = c(s, s')/\sum s'' c(s, s'')
\]

\[
\lambda(s) = c(s)/\sum s' c(s')
\]

- \( C(s, s') \) is the count of the number of times \( s' \) follows \( s \) in the training data. An element of the matrix \( A \), say \( A[s, s'] \) can be interpreted as the probability of transitioning from state \( s \) to \( s' \) in one step. Similarly an element of \( A^*A \) will denote the probability of transitioning from one state to another in two steps, and so on.
Markov Chain Models for Link Prediction (2)

• Given the "link history" of the user $L(t-k), L(t-k+1).... L(t-1)$, we can represent each link as a vector with a probability 1 at that state for that time (denoted by $i(t-k), i(t-k+1)...i(t-1)$). The Markov Chain models estimation of the probability of being in a state at time 't' is:
  \[ s(t) = i(t-1)A \]

• A proposed variant of the Markov process to accommodate weighting of more than one history state is:
  \[ s(t) = a_0 i(t-1)A + a_1 i(t-2)A^2 + ..... \]
  \[ s(t) = \max(a_0 i(t-1)A, a_1 i(t-2)A^2, ..... ) \]
Applications(1)

• **Web Server HTTP Request Prediction**
  – The client sends a request to the web server (or proxy) which uses the HTTP probabilistic link prediction module
  – The server uses the Markov chain model in an adaptive mode, updating the transition matrix using the sequence of requests that arrive at the web server to predict the links that this client may be interested in.
Applications(2)

- **Adaptive Web Navigation**
  - Link prediction is used to build a navigation agent which suggests (to the user) which other sites/links would be of interest to the user based on the statistics of previous visits (either by this particular user or a collection of users').
  - The predicted link doesn't strictly have to be a link present in the web page currently being viewed.
  - If the link modeling is user-specific then the link predictor module can be resident at the client side rather than the server side.
Applications(3)

• **Tour Generation**
  – The tour generator module is given as input the starting URL (e.g. the current document the user is browsing).
  – The tour module generates a sequence of states (or URLs) using the Markov Chain process. This is returned and displayed to the client as a tour.