Information Retrieval

Lecture 11 - Link analysis

Seminar für Sprachwissenschaft

International Studies in Computational Linguistics

Wintersemester 2007

Introduction

- Link analysis: using hyperlinks for ranking web search results
- Link analysis is only one of the factors used by search engines to compute a score on a given query
- Note that counting in-links is not enough (*cf* spam links)
- Link analysis is comparable to citation analysis (authority of a paper ≡ amount of citations)

Overview

Recall: web as a graph

. . .
↓
Page A
text 1
text 2
→ Page B ← . . .
↓
Page C
↑
. . .

Recall: web as a graph (continued)

- 2 observations:
  - (a) anchor text pointing to a page B is a good description of page B
  - (b) hyperlink from page A to page B is an endorsement of page B
- (a) helps for indexing pages that do not contain the terms people usually use to refer to them, and also for indexing images and other specific content
- (a) can be used conjoinly with the analysis of a window of terms surrounding the anchor
- (b) suggests that observing the distribution of links within the web can help finding the most relevant pages

Overview

About PageRank

- PageRank: scoring measure based only on the link structure of web pages
- Every node in the web graph is given a score between 0 and 1, depending on its in and out-links
- Given a query, a search engine combines the PageRank score with other values to compute the ranked retrieval (*e.g.* cosine similarity, relevance feedback, etc.)

About PageRank (continued)

- Underlying idea: computing a score reflecting the "visitability" of a page
  - When surfing the web, a user may visit some pages more often than others (*e.g.* more in-links)
  - Pages that are often visited are more likely to contain relevant information
  - When a dead-end page is reached, the user may teleport (*e.g.* type an address in the browser)
- NB: the teleport operation consists of a uniform choice at random within the nodes of the web graph

Assigning a PageRank score

- Based on a traversal of the web graph:
  1. When a page has no out-links, the user teleports
  2. When a page has out-links, the user may teleport with a probability α ($0 \leq \alpha \leq 1$)
- In the second case, the probability for the user to click on an out-link is $1 - \alpha$ (α is generally set to 0.1)
- When the surfer follows this schema for a certain time, he visits each node v a fixed fraction of time $\pi(v)$ that depends on both the structure of the graph and α

  $\pi(v) \equiv$ PageRank of v

Overview

Markov chains

- Discrete stochastic process, corresponding to a list of steps at which a random choice is make
- Can be characterized by an $N \times N$ transition probability matrix P, where:

$$0 \leq P_{i,j(1 \leq i,j \leq N)} \leq 1 \qquad \sum_{j=1}^{N} P_{ij} = 1 \; \forall i \in [1..N]$$

- P_{ij} gives the probability, being at time t in step i, to be in step j at time $t+1$ (called transition probability)
- Note that there is no memory (*i.e.* the probability only depends on the current step, not the previous ones)

Markov chains (continued)

- Property of such stochastic matrices: they have a principal left eigenvector for the largest eigenvalue 1
- Recall: an vector $\vec{v}$ is an eigenvector for a matrix M iff

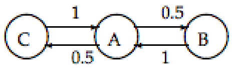
$$M.\vec{v} = \lambda.\vec{v}$$

  (λ is the eigenvalue for the eigenvector $\vec{v}$)
- An eigenmatrix can be decomposed as follows:

$$M = Q \Lambda Q^{-1}$$

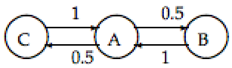
  where:
  - Q is such that Q_i is the i$^{\text{th}}$ eigenvector
  - Λ is a diagonal matrix, and Λ_{ii} is the i$^{\text{th}}$ eigenvalue
- Q_1 is the left eigenvector and Λ_{11} the corresponding eigenvalue

Markov chains: example

- 

  Example from (Manning et al., 2008)

Markov chains: example

- 

  Example from (Manning et al., 2008)
-
$$\begin{array}{c} \\ A: \\ B: \\ C: \end{array} \begin{array}{c} \begin{array}{ccc} A: & B: & C: \end{array} \\ \begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \end{array}$$

Probability vectors

- A probibility vector $\vec{v}$ is such that:
$$v_{i(1 \leq i \leq N)} \in [0,1] \quad \wedge \quad \sum_{i=1}^{N} v_i = 1$$
- A probability vector $\vec{v} = (v_1 \dots v_N)$ defines a state in the chain
  $\vec{v} = (\overset{1}{0} \dots \overset{i}{1} \dots \overset{N}{0})$ refers to state i
- More generally, the component v_i of a probability vector defines the probability to be in state i
- If $\vec{v}$ is the probability of the current step, the probability of the next step is $\vec{v}.P$

Ergodic Markov chains

- A Markov chain is said to be <u>ergodic</u> when:
$$\exists T_0 \in \mathcal{R}^+ \text{ such that } \forall i, j \in [1..N]$$
  $\forall t > T_0$ the probability to be in state j is ≥ 0
- To be ergodic, a Markov chain needs to have 2 properties:
  - irreductability: for all states i, j there is a sequence of transitions from i to j with non-zero probability
  - aperiodicity: no partition of the states into sets, from which only cycles are defined

Property of an ergodic Markov chain

- For any ergodic Markov chain, there exists a unique probability vector $\vec{\pi}$ which is the principal left eigenvector of the probability matrix P, and such that, if we note $\eta(i,t)$ the number of visits to the state j in t steps:

$$\lim_{t \to +\infty} \frac{\eta(i,t)}{t} = \pi_i > 0$$

- π_i is the steady-state probability for state i

Overview

Recall: web as a graph

PageRank
- About PageRank
- Markov chains
- Markov chains and the web
- Computing the PageRank score

Topic-specific PageRank

Hubs and authorities

Markov chains and the web

- A random web surf can be seen as a Markov chain
- Considering the adjacency matrix A such that, $\forall i,j \in [1..N]$:

$$A_{ij} = \begin{cases} 1 & \text{iff there is a link from page } i \text{ to page } j \\ 0 & \textit{otherwise} \end{cases}$$

- The $N \times N$ probability matrix P of a web surf is built using the following algorithm:
  1) each 1 in A is divided by the number of 1 in its row
  2) the resulting matrix is multiplied by $1 - \alpha$
  3) $\frac{\alpha}{N}$ is added to every entry of the resulting matrix (for teleport probability)

  $\rightarrow$ the resulting matrix is P

Markov chains and the web: example

- Considering the following adjacency matrix:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

- Compute the probability matrix, for a teleport operation of probability $\alpha = 0.5$

Example from (Manning et al., 2008)

Markov chains and the web (continued)

- In case of a "long" traversal of the web (*i.e.* of the Markov chain), each state is visited at a different frequency
- If we consider the Markov chain representing the web to be ergodic:

  $\rightarrow$ this frequency of visits converges to a fixed steady-state quantity

  $\rightarrow$ the PageRank score

Outline

Computing the PageRank score

- One way to compute the PageRank score is to use the power iteration method, based on the following remark:

  if $\vec{\pi}$ is a steady-state distribution and P the probability matrix, we have:

  $$\vec{\pi} = \lambda.\vec{\pi}$$

  In other terms, $\vec{\pi}$ is the left eigenvector of P, whose eigenvalue is 1:

  $$\vec{\pi} = 1.\vec{\pi}$$

Computing the PageRank score (continued)

- Wherever we start, after some iterations, we reach the steady state $\vec{\pi}$
- If the initial probability vector (initial step) is $\vec{x}$,
  - after one step, we are in $\vec{x}.P$
  - after two steps, we are in $\vec{x}.P^2$
  - and so on.
- For a "large" k, $\vec{x}.P^k = \vec{\pi}$

Computing the PageRank score: example

- If we consider the following probability matrix:

$$P = \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}$$

- If the surf begins at state 1, compute the 3 first transitions
- Note that the steady-state vector is reached after several iterations and is:

$$\vec{\pi} = (5/18 \;\; 4/9 \;\; 5/18)$$

Example from (Manning et al., 2008)

Remarks about the PageRank score

- The PageRank score is independent from any query
  $\rightarrow$ static quality measure of a web page
- Thus the Pagerank score is pre-processed by (i) building the probability matrix for a graph of web pages, and
  (ii) computing its steady state (iteratively or not)
- At run-time, the query is processed to retrieve a given amount of pages, which are then ranked using the query-independent PageRank score
- Note hat the Markov model presented here do not take the back button into account

Outline

Topic-specific PageRank

- Relies on the fact that the teleport operation cannot considered to be chosen at random uniformly
- Some pages are more likely to be entered in the browser than others
- Idea: the teleport operation is chosen at random uniformly within a given topic
- These topics are defines via either:
  - a manually built directory of pages (*e.g.* open directory project `http://www.dmoz.org`)
  - an automatic text classification algorithm

Topic-specific PageRank (continued)

- Use a personalized PageRank score
- Idea: Approximation of the interests of the search user as a linear combination of a small number of topics
- Each elementary PageRank score (one per topic) is pre-computed considering a teleport operation within this given topic
- Then, the personalized PageRank for a given users is computed on the fly as a linear combination of the elementary PageRank scores

  Example: user interested mainly in sports (60%) and health (40%)

$$\vec{\pi} = 0.6\vec{\pi_{sport}} + 0.4\vec{\pi_{health}}$$

Outline

Hubs and authorities

- Underlying idea: there are 2 main kinds of useful pages for broad-topic searches
  - authoritative sources of information or authorities (*e.g.* medical research institute)
  - hand-compiled lists of authoritative sources or hubs (*e.g.* association promoting health-care)
- Basic property of such pages:
  - good hubs points to many good authorities
  - good authorities are pointed by many hubs
- In this context, given a query, web pages will be given 2 scores: a "hub score" and a "authority score"

Hubs and authorities (continued)

Iterative computation of these scores:

- Starting point: a subset S of "good" hubs and authorities
  All nodes v are given the scores

$$h(v) = a(v) = 1$$

- Iteration:

$$h(v) \leftarrow \sum_{v \to y} a(y) \qquad a(v) \leftarrow \sum_{y \to v} h(y)$$

- Using the adjacency matrix A, this can ber expressed as:

$$\vec{h} = A.\vec{a} \qquad \vec{a} = A^T.\vec{h}$$

- Thus:

$$\vec{h} = A.A^T\vec{h} \qquad \vec{a} = A^T.A\vec{a}$$

Hubs and authorities (continued)

- To sum up:
  1) gather a subset of web pages
  2) compute the adjacency matrix A, $A.A^T$ and $A^T.A$
  3) compute the left eigenvectors of $A.A^T$ and $A^T.A$ which are respectively $\vec{h}$ and $\vec{a}$
- How to select the starting subset ?
  a) given a query, use a text index to get all pages containing the terms $\Rightarrow$ root set
  b) add all pages that either point to or are pointed by pages of the root set $\Rightarrow$ base set

Hubs and authorities (continued)

- Method known as *Hyperlink-Induced Topic Search* (HITS)
- Top hubs and authorities include other languages than those of the query (cross-language retrieval)
- 200 pages are enough for the root set
- 5 iterations are usually enough to compute the top hubs and authorities
- We are more interested in relative scores than absolute ones, thus during iterations $\vec{a}$ and $\vec{h}$ can be scaled down
- In practice, addictive updates are used rather than matrix products (time complexity)
- Main issues:
  - off-topic authorities (*e.g.* super-topic)
  - bias via affiliated web pages

Conclusion

- Link analysis is used to guide the crawling of the web and to give a static score to a web page
- This static score is one of the components of the final score a page gets for a given query
- The PageRank algorithm relies on Markov chains to compute a score corresponding to a steady state in terms of frequency of visits of a page during a web surf
- The HITS algorithm is used for broad-topic searches and computes a score from the idea that reliable hubs connect reliable authorities

References

- C. Manning, P. Raghavan and H. Schütze
  Introduction to Information Retrieval
  `http://nlp.stanford.edu/IR-book/pdf/chapter21-linkanalysis.pdf`
- Lawrence Page and Sergey Brin
  The PageRank Citation Ranking: Bringing Order to the Web (1998)
  `http://citeseer.ist.psu.edu/page98pagerank.html`
- Jon Kleinberg
  Authoritative Sources in a Hyperlinked Environment (1999)
  `http://citeseer.ist.psu.edu/kleinberg99authoritative.html`