Introduction to SALSA (Stochastic Approach for Link-Structure Analysis)
• A fundamental problem in information retrieval is ranking.

• Web search engines have a number of additional features at their disposal, including the hyperlinks leading from one web page to another.

• A hyperlink can be viewed as an endorsement by a web page’s author of another web page.
• Link-based ranking algorithms can be broadly grouped into two classes:
  – Query independent algorithms that estimate the quality of a web page, and
  – Query-dependent ones that estimate its relevance to a particular query.

• Recent research has shown that query-dependent link-based ranking algorithms (notably, the SALSA algorithm) are substantially more effective than well-known query-independent ones such as PageRank.
In the mid-1990s, Jon Kleinberg proposed an algorithm called *Hypertext-Induced Topic Search* or HITS for short.

HITS is a query-dependent algorithm: It views the documents in the result set as a set of nodes in the web graph; it adds some nodes in the immediate neighborhood in the graph to form a *base set*, it projects the base set onto the full web graph to form a neighborhood graph, and finally it computes two scores, a *hub* score and an *authority* score, for each node in the neighborhood graph.

The authority score estimates how relevant a page is to the query that produced the result set; the hub score estimates whether a page contains valuable links to authoritative pages.

Authority and hub scores mutually enforce each other.
• SALSA is a variation of Kleinberg’s algorithm.
• takes a result set R as input, and constructs a neighborhood graph from R in precisely the same way as HITS.
• Similarly, it computes an authority and a hub score for each vertex in the neighborhood graph, and these scores can be viewed as the principal eigenvectors of two matrices.
• However, instead of using the straight adjacency matrix that HITS uses, SALSA weights the entries according to their in and out-degrees.
• The approach is based upon the theory of Markov chains, and relies on the stochastic properties of random walks performed on our collection of pages.

• The input to our scheme consists of a collection of pages $C$ which is built around a topic $t$.

• Intuition suggests that authoritative pages on topic $t$ should be visible from many pages in the subgraph induced by $C$. Thus, a random walk on this subgraph will visit $t$-authorities with high probability.
Formal Definition of SALSA

• Let us build a bipartite undirected graph $G = (V_h, V_a, E)$ from our page collection and its link-structure:
  – $V_h = \{sh|S \in C \text{ and out-degree}(s) > 0\}$ (the hub side of $G$).
  – $V_a = \{sa|S \in C \text{ and in-degree}(s) > 0\}$ (the authority side of $G$).
  – $E = \{(sh, ra)|s \Rightarrow r \text{ in } C\}$.

• Each non-isolated page $s \in C$ is represented in $G$ by one or both of the nodes $sh$ and $sa$. Each WWW link $s \Rightarrow r$ is represented by an undirected edge connecting $sh$ and $ra$.

• On this bipartite graph we will perform two distinct random walks. Each walk will only visit nodes from one of the two sides of the graph.
We will examine the two different Markov chains which correspond to these random walks:

- the chain of the visits to the authority side
- the chain of the visits to the hub side

The hub matrix is defined as:

\[ h_{i,j} = \sum_{\{k \mid (k_h, i_a), (k_h, j_a) \in G\}} \left( \frac{1}{\deg(i_a)} \right) \left( \frac{1}{\deg(k_h)} \right) \]
The authority matrix is defined as:

\[
\begin{align*}
    a_{i,j} = & \sum_{\{k \mid (k_h, i_a), (k_h, j_a) \in G\}} \left( \frac{1}{\text{deg}(i_a)} \right) \cdot \left( \frac{1}{\text{deg}(k_h)} \right).
\end{align*}
\]

A positive transition probability \( a(i, j) > 0 \) implies that a certain page \( k \) points to both pages \( i \) and \( j \), and hence page \( j \) is reachable from page \( i \) by two steps: retracting along the link \( k \rightarrow i \) and then following the link \( k \rightarrow j \).
• Let $W$ be the adjacency matrix of the directed graph defined by and its link structure.

• Denote by $Wr$ the matrix which results by dividing each nonzero entry of $W$ by the sum of the entries in its row, and by $Wc$ the matrix which results by dividing each nonzero element of $W$ by the sum of the entries in its column.

• $H$ consists of the nonzero rows and columns of $W_rW_c^T$, and $A$ consists of the nonzero rows and columns of $W_c^TW_r$. 