Web Intelligence

Web Communities and Dissemination of Information and Culture on the www

The HITS Algorithm: Web Communities

Google's PageRank gives a score to every page, in order to help with relevance and usefulness in search.

HITS (Hyperlink-Induced Topic Search) is an alternative method, which tries to find the key pages for specific **web communities**.

HITS focuses on finding **authorities** (pages which many inlinks) and **hubs** (pages with many outlinks) that are relevant to specific topics (such as may be gleaned from a search query).

Authorities and Hubs

Suppose Rq is a set of pages that have been retrieved by a search engine for a specific query q.

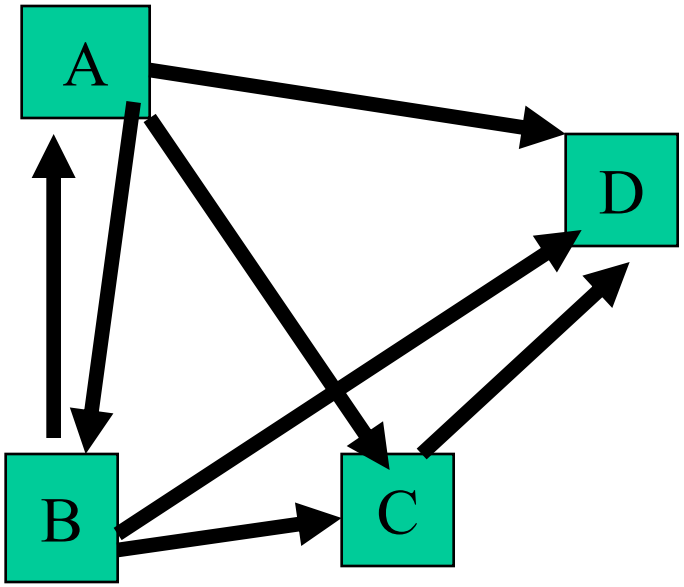
Let Ai be the authority score for page i, and let Hi be the hub score for page i. We can initialise these at 1 for every page, and then iterate the following two equations until the numbers settle down:

$$Ai = \sum_{\text{pages } x \text{ that link to } i} Hx$$

$$Hi = \sum_{\text{pages } x \text{ that } i \text{ points to}} Ax$$

Authorities and Hubs example

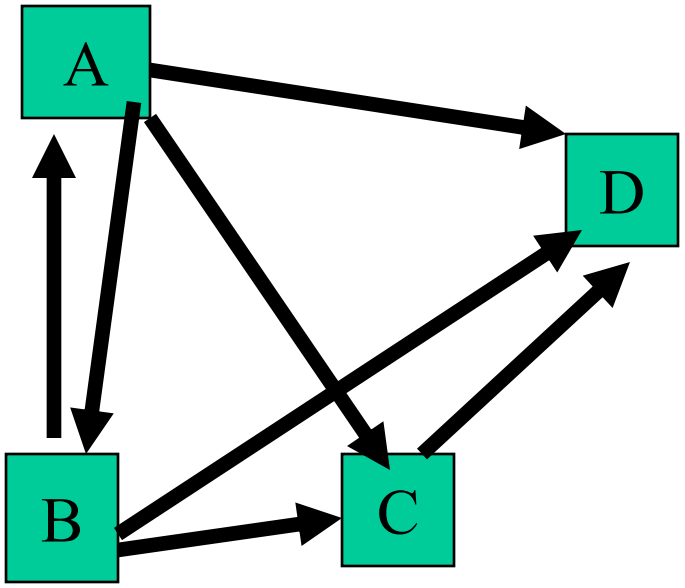
Initially Ha = Hb = Hc = Hd =1



1. Aa = Hb = 1; Ab = Ha = 1; Ac = Ha + Hb = 2; Ad = Ha + Hb + Hc = 3
   Normalise: Aa = 0.143 ; Ab = 0.143; Ac = 0.286; Ad = 0.429
   Ha = Ab + Ac + Ad = 0.858; Hb = Aa + Ac + Ad = 0.858; Hc = 0.429; Hd = 0
   Normalise: Ha = 0.4; Hb = 0.4; Hc = 0.2; Hd = 0

Authorities and Hubs example

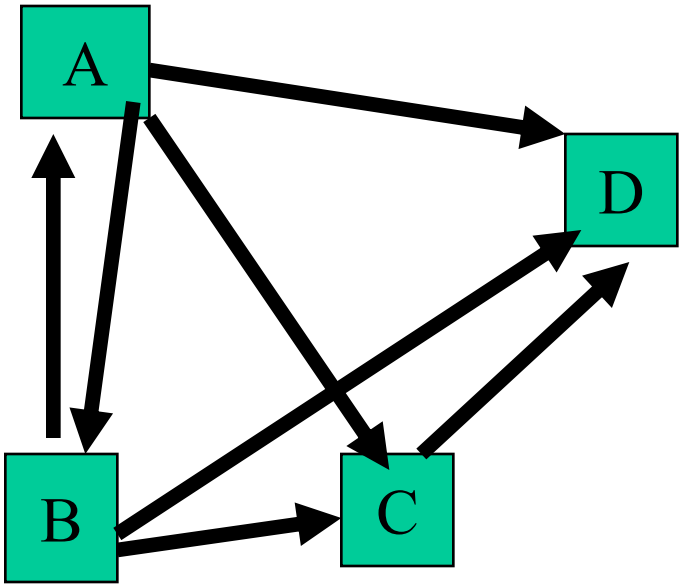
Initially Ha = Hb = Hc = Hd =1



2. Aa = Hb = 0.4; Ab = Ha = 0.4; Ac = Ha + Hb = 0.8; Ad = Ha + Hb + Hc = 1
   Normalise: Aa = 0.154 ; Ab = 0.154; Ac = 0.308; Ad = 0.386
   Ha = Ab + Ac + Ad = 0.848; Hb = Aa + Ac + Ad = 0.848; Hc = 0.386; Hd = 0
   Normalise: Ha = 0.356; Hb = 0.356; Hc = 0.288; Hd = 0

Authorities and Hubs example

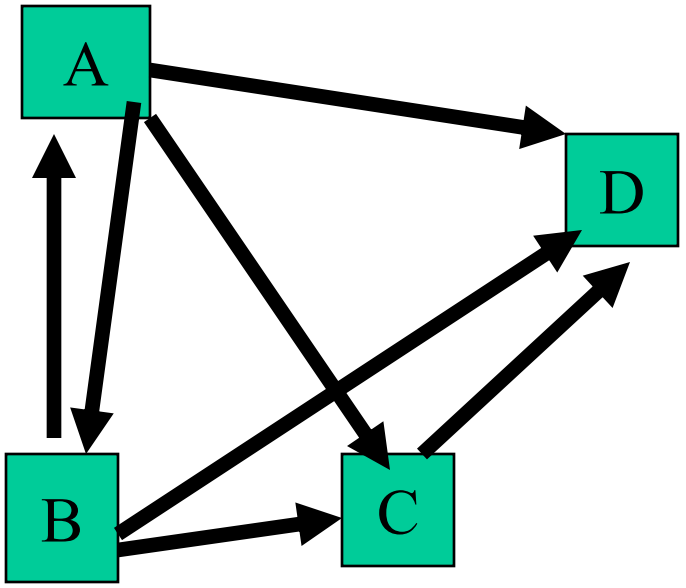
Initially Ha = Hb = Hc = Hd =1



3. Aa = Hb = 0.356; Ab = Ha = 0.356; Ac = Ha + Hb = 0.712; Ad = Ha+Hb+Hc = 1
   Normalise: Aa = 0.146 ; Ab = 0.146; Ac = 0.292; Ad = 0.416
   Ha = Ab + Ac + Ad = 0.854; Hb = Aa + Ac + Ad = 0.854; Hc = 0.416; Hd = 0
   Normalise: Ha = 0.402; Hb = 0.402; Hc = 0.196; Hd = 0

Authorities and Hubs example

Initially Ha = Hb = Hc = Hd =1



4. Aa = Hb = 0.402; Ab = Ha = 0.402; Ac = Ha + Hb = 0.804; Ad = Ha+Hb+Hc = 1
   Normalise: Aa = 0.154 ; Ab = 0.154; Ac = 0.308; Ad = 0.384
   Ha = Ab + Ac + Ad = 0.846; Hb = Aa + Ac + Ad = 0.846; Hc = 0.384; Hd = 0
   Normalise: Ha = 0.408; Hb = 0.408; Hc = 0.184; Hd = 0

… eventually the numbers converge

Authorities and Hubs exhibit *mutually reinforcing relationships*.

- A good hub points to many good authorities
- A good authority is pointed to by many good hubs.

(as is also true with PageRank …) the calculation is done in a different way. This is indicated in the HITS algorithm pseudocode on the next slide.

This alg says how HITS responds to a query, *q*

Contrast this with how google deals with *q*.

The HITS Algorithm

1. Get the *r* highest ranked pages for query *q;* call the pages Rq
2. Expand these to set Sq, containing all pages pointed to by pages in Rq, and add up to *d* pages that point to pages in Rq.
3. Consider the link graph of Sq, *G*. There are **transverse** links (between pages in Sq that have different domain names), and **intrinsic** links (between pages with the same domain name). Delete all intrinsic links of *G*
4. Obtain a ranked list of **authorities** in *G*. This can be done by the simple repeated iteration of authority scores and hubs scores. But it is done in practice by:
   - Form the adjacency matrix of *G*, A, and its transpose A^T
   - Find the normalised principal eigenvector e of A^TA
   - Values in this eigenvector correspond to authority scores.

(reasonable parameter values: $r = 200$; $d = 50$ – leads to around 1000—5000 pages in Sq)

A Problem with HITS

Examinable Reading: the Nomura et al paper, sections 1, 2 and 3.

Understand the problem.

Main point: any technique for deciding on the importance of a web page can be misled, either deliberately or not, by certain link structures. For example, how might you deceive the PageRank method into thinking that your www page was important?

Cultural Dissemination: Axelrod's Model

Axelrod formulated a simple and very influential **model of cultural dissemination**. That is: how ideas, traits, characteristics, fashions, etc …, spread in communities.

This model (and culture dissemination models in general) help us understand the factors that lead to:

- *Globalisation*: where an entire community becomes very similar in some aspect (e.g. everyone using google? everyone using English as the language of science?)
- *Polarization*: for a particular aspect, the community divides into two distinct choices – e.g. Windows users and Mac users
- *Differentiation*: more than two stable sub-communities: e.g. the presence of different religious groups.

Assumptions in Cultural Models

The two key bases in a cultural model are:

- People like to change, to become a little more like the people in their own social group. E.g. wear similar clothes, go to similar restaurants, adopt similar views.

- People are more likely to be influenced by those who are already similar to them. E.g. a Norwegian goatherder will consider buying boots that are like his respected neighbour's boots, but not the Australian prime minister's boots.

<u>These assumptions are demonstrably true</u>. So, why doesn't everyone eventually become the same? How long does it take for globalisation to occur for a particular aspect? These and other questions are explored by using cultural models.

Axelrod's model of cultural dissemination

Individuals are placed on a spatial grid – although any spatial structure can be used.

Each individual has F features (e.g. religion, fashion, diet, …), and each feature has q possible values.

The feature vectors are initially random (or otherwise, depending on what experiment you want to do).

The model typically runs as follows:

1. **A random individual *a* is chosen, and a random neighbour of that individual, *b*, is chosen**
2. **If *a* and *b* have x features in common ($1 < x < F$), then *a* will change to match another one of *b*'s features, with probability q/F.**

Simple Example

PC, Jedi, Car, C

Mac, Jedi, Bike, C

PC, Jedi, Car, Java

Mac, Jedi, Car, Java

Mac, Sith, Bike, C

Mac, Jedi, Car, Java

Mac, Jedi, Car, Java

PC, Sith, Bike, C

Mac, Jedi, Bike, C

PC, Jedi, Bike, Java

PC, Sith, Bike, C

Mac, Jedi, Bike, Java

Simple Example

PC, Jedi, Car, C	Mac, Sith, Bike, C	Mac, Jedi, Bike, C
Mac, Jedi, Bike, C	Mac, Jedi, Car, Java	PC, Jedi, Bike, Java
PC, Jedi, Car, Java	Mac, Jedi, Car, Java	PC, Sith, Bike, C
Mac, Jedi, Car, Java	PC, Sith, Bike, C	Mac, Jedi, Bike, Java

Choose a random individual (red) and a neigbour (blue)

Simple Example

PC, Jedi, Car, C	Mac, Sith, Bike, C	Mac, Jedi, Bike, C
Mac, Jedi, Bike, C	PC, Jedi, Car, Java	PC, Jedi, Bike, Java
PC, Jedi, Car, Java	Mac, Jedi, Car, Java	PC, Sith, Bike, C
Mac, Jedi, Car, Java	PC, Sith, Bike, C	Mac, Jedi, Bike, Java

This individual changes to be a bit closer to neighbour

Simple Example

PC, Jedi, Car, C	Mac, Sith, Bike, C	Mac, Jedi, Bike, C
Mac, Jedi, Bike, C	PC, Jedi, Car, Java	PC, Jedi, Bike, Java
PC, Jedi, Car, Java	PC, Jedi, Car, Java	PC, Sith, Bike, C
Mac, Sith, Car, Java	PC, Sith, Bike, C	Mac, Jedi, Bike, Java

Another random individual and a neighbour

Simple Example

PC, Jedi, Car, C	Mac, Sith, Bike, C	Mac, Sith, Bike, C
Mac, Jedi, Bike, Java	PC, Jedi, Car, Java	PC, Jedi, Bike, Java
PC, Jedi, Car, Java	PC, Jedi, Car, Java	PC, Sith, Bike, C
Mac, Sith, Car, Java	PC, Sith, Bike, C	Mac, Jedi, Bike, Java

Again a random individual is chosen, and a neighbour, but these two are already the same, so no change.

Simple Example

PC, Jedi, Car, C	Mac, Sith, Bike, C	Mac, Jedi, Bike, C
Mac, Jedi, Bike, Java	PC, Jedi, Car, Java	PC, Jedi, Bike, Java
PC, Jedi, Car, Java	PC, Jedi, Car, Java	PC, Sith, Bike, C
Mac, Sith, Car, Java	PC, Sith, Bike, C	Mac, Jedi, Bike, Java

No change – they are too different, so the individual is not influenced by the neighbour

Eventually, something like this happens …

Mac, Sith, Bike, C	Mac, Sith, Bike, C	Mac, Sith, Bike, C
Mac, Sith, Bike, C	PC, Jedi, Car, Java	PC, Jedi, Car, Java
PC, Jedi, Car, Java	PC, Jedi, Car, Java	PC, Jedi, Car, Java
Mac, Sith, Bike, C	Mac, Sith, Bike, C	Mac, Sith, Bike, C

The community has **polarized** into two distinct types.
Alternatively, it may have ***globalized***, or it may have **differentiated**.

What happens depends on subtleties of the parameters in context.
I.e. the degree of difference thresholds within which individuals will be influenced by neighbours; the size of the neighbourhood, and so on.

Proper examples: evolution of `cultural domains'

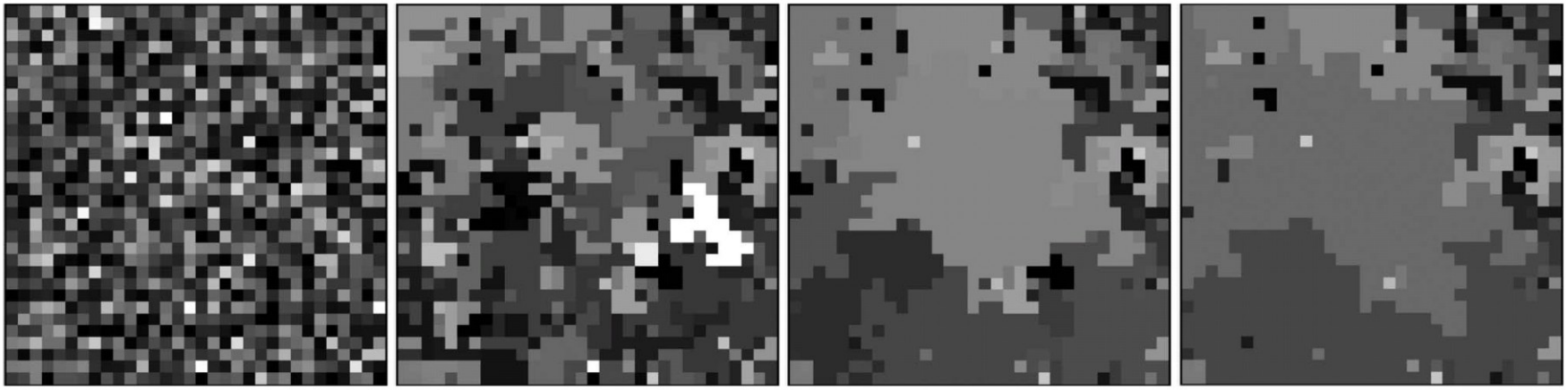


Figure 4. Dynamical evolution. For a system size $N = 32 \times 32$ with parameter values $F = 3$ and $q = 10$, snapshots of the time evolution of Axelrod's model from random initial conditions at times $t = 0$, 1,000, 3,000, and 6,807 show the emergence of cultural domains. At time $t = 6{,}807$, the dynamics stop and the configuration freezes.

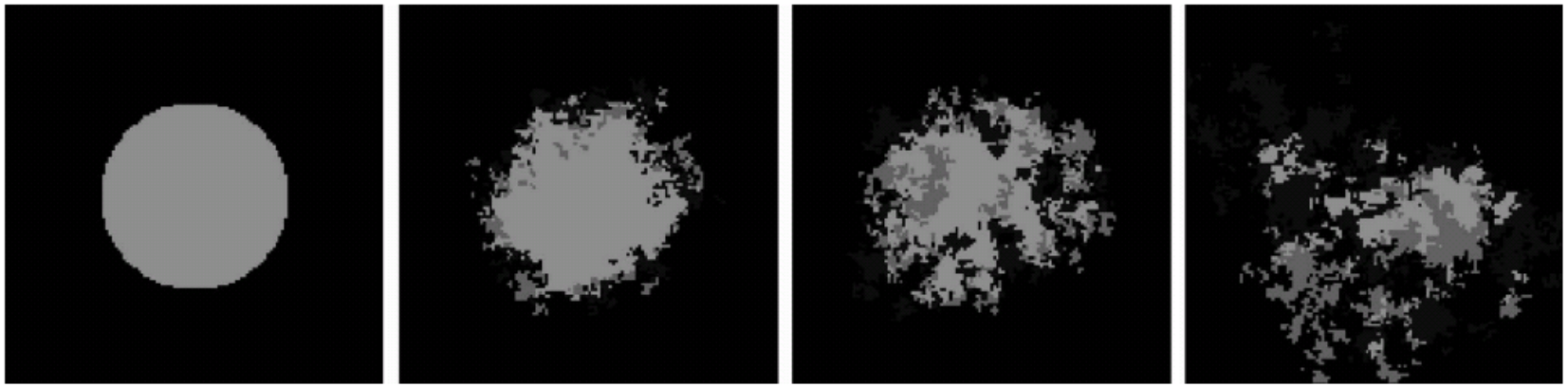


Figure 3. Axelrod's model. For a system size $N = 128 \times 128$ with parameter values $F = 3$ and $q = 15$, different colors indicate different cultural states in snapshots of the model's evolution at times $t = 0$, 114, 272, and 1,331.

Interesting transitions

Globalisation

This axis shows the size of stable communities that emerge

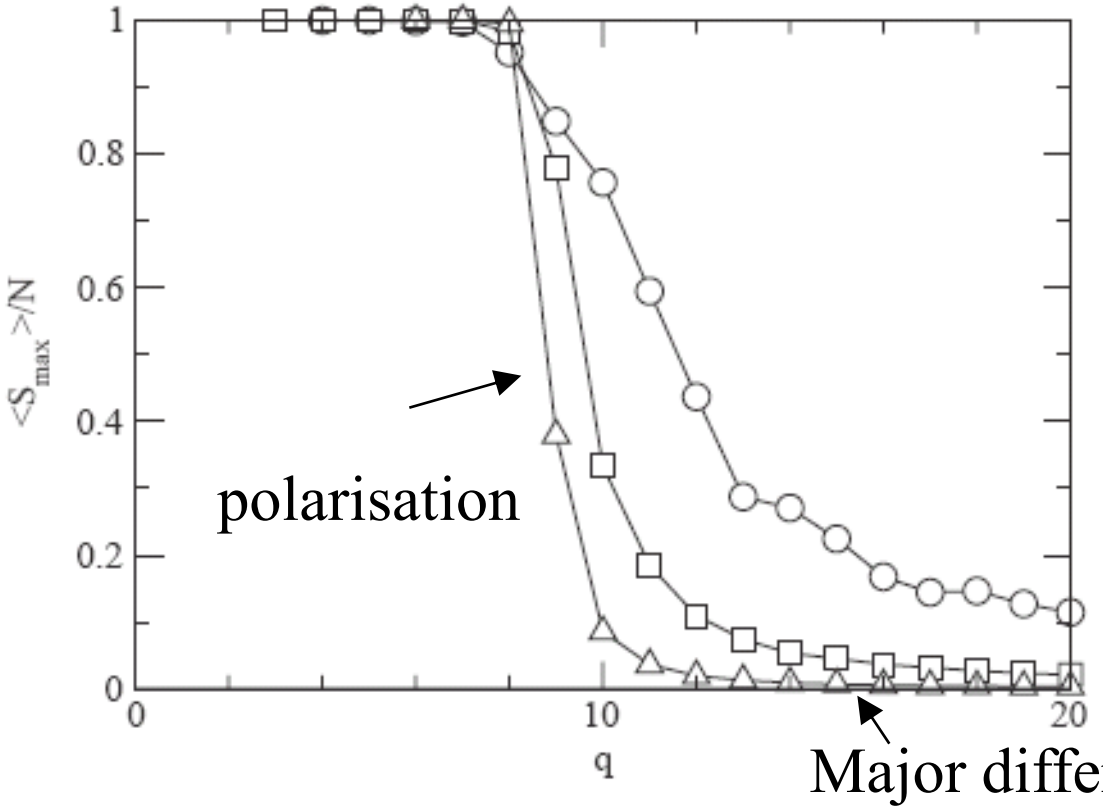


Fig. 4. The average $\langle S_{\max} \rangle / N$ in one-dimensional lattices as a function of q for $N = 100$ (circles), 1000 (squares), 10 000 (triangles) agents. Each plotted value is an average over 100 runs with independent initial conditions. Number of features $F = 10$.

Globalisation apparent when few choices for a feature – polarisation more common in small communities?

- Lots of research starting to be done on spread of ideas and culture on the WWW – using Axelrod-style models on web graphs.
- Think about examples of polarisation and globalisation on the www that you think have happened, or are likely to happen.
- The papers I got the last three figures from are on the www as recommended reading.

- The end