





	WED SEARCH ENGINES (WSES) 01 THE SECOND GENERATION	
	(from 1998 onwards)	
	Identify relevance with topic-relateness and authoritativeness	
	Independent by the particular format of the Web site	
_	Relevance computation is more selective	
Ц	Ranking Schemes algorithms which compute	
	authoritativeness exploiting the hyperlink structure of the	
п	Web The Web can be seen as a network of recommendations a	
	social network. Social networks analysis has been applied in	
	many contexts in the past, including epidemiology,	
Dip.	di Matematica F. Aiolli - Sistemi Informativi 7	
Pura	Pura ed Applicata 2006/2007	
S	nam W/ah Sites	
S	pam Web Sites	
S	pam Web Sites (SWSs) are Web pages designed to	
S	pam Web Sites Spam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for	
S	pam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for commercial purposes	
S	 pam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for commercial purposes First Generation WSEs Including deceptive self-description in the HTML META tag 	
S	 pam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for commercial purposes First Generation WSEs Including deceptive self-description in the HTML META tag Including "invisible words" (i.e. displayed in the same color as 	
S	 pam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for commercial purposes First Generation WSEs Including deceptive self-description in the HTML META tag Including "invisible words" (i.e. displayed in the same color as the background) or words typeset in tiny fonts, in order to deceive tfidf-based ranking schemes 	
S	 pam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for commercial purposes First Generation WSEs Including deceptive self-description in the HTML META tag Including "invisible words" (i.e. displayed in the same color as the background) or words typeset in tiny fonts, in order to deceive tfidf-based ranking schemes Second Generation WSEs 	
S	 pam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for commercial purposes First Generation WSEs Including deceptive self-description in the HTML META tag Including "invisible words" (i.e. displayed in the same color as the background) or words typeset in tiny fonts, in order to deceive tfidf-based ranking schemes Second Generation WSEs LRSs would seem to be more robust, since SWSs are not 	
S	 pam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for commercial purposes First Generation WSEs Including deceptive self-description in the HTML META tag Including "invisible words" (i.e. displayed in the same color as the background) or words typeset in tiny fonts, in order to deceive tfidf-based ranking schemes Second Generation WSEs LRSs would seem to be more robust, since SWSs are not authoritative, but naive LRSs may be fooled by artificially conferring authority onto SWSs 	
S	 pam Web Sites (SWSs) are Web pages designed to manipulate WSE ranking schemes, generally for commercial purposes First Generation WSEs Including deceptive self-description in the HTML META tag Including "invisible words" (i.e. displayed in the same color as the background) or words typeset in tiny fonts, in order to deceive tfidf-based ranking schemes 	

Dip. di Matematica
Pura ed Applicata



Dip.	di I	Matematica	
Pura	ed	Applicata	

LRSs



Adjacency Matrix

The input to any LRS is thus a |BS|×|BS| adjacency matrix W such that

W[i,j]=1 iif there is a hyperlink from page p_i to p_j

- □ The output of any LRS is a vector $a=[a_1,..,a_{|BS|}]$ where a_i is the authoritativeness of page p_i
- Backward Neighbors, B(i)={p_j | W[j,i]=1}
 Forward Neighbors, F(i)={p_j | W[i,j]=1}

The InDegree Algorithm The InDegree algorithm [Marchiori97], consists in identifying the authoritativeness a; of a page p; with the in-degree of p_i , i.e. |B(i)|It corresponds to ranking Web pages according to their 'popularity' ('visibility') In matric notation $a = W^{T} \mathbf{1}$ Main weakness: only the quantity of backward links, and not their quality, matters □ It can fooled easily by SWSs. To promote a page p_s, they only need to set up lots of dummy pages $p_1, ..., p_k$, containing pointers to p. Not used in any current-day WSE F. Aiolli - Sistemi Informativi 13 Dip. di Matematica Pura ed Applicata 2006/2007

Pagerank scoring

- Imagine a browser doing a random walk on web pages:
 - Start at a random page



- At each step, go out of the current page along one of the links on that page, equiprobably
- "In the steady state" each page has a long-term visit rate - use this as the page's score.







F. Aiolli - Sistemi Informativi
2006/2007

Dip. di Matematica

Pura ed Applicata