**Combating Web Spam with TrustRank---** Zoltán Gyöngyi, Hector Garcia-Molina and Jan Pedersen

이 논문은 good page를 스팸으로부터 분리하는 semi-automatically한 방식을 소개한다. 먼저 전문가가 고른 seed page set을 정하고 웹의 링크 구조를 이용하여 이외의 다른 좋은 페이즈들을 찾아낸다. seed selection을 하는 방식과 good page를 찾아내는 방식이 설명되어 있다. web spam은 웹상에서 서치 엔진을 misleading할 의도로 만들어진 하이퍼링크 페이지를 말한다. 스팸의 방식으로는 홈페이지에 수 천개의 키워드를 추가해서 검색되게 하거나, 특정 타겟 페이지를 가리키는 수많은 가짜 웹 페이지를 만들어 incoming link의 숫자를 증가시킨 후 검색되게 만드는 것 등이 있다. 이메일 스팸같이 웹 페이지가 스팸인지 아닌지 판단하는 일은 주관적이다. 다른 웹 페이지와 반복적으로 링크를 갖는 웹 사이트의 클러스터가 있다고 생각할 때, 이 링크들은 사이트 간에 유용한 관계를 표현할 수도 있고 다른 페이지의 랭크를 높이기 위한 목적으로 사용 됬을 수도 있다. 때문에 이 두 가지 경우를 구별하는 일은 어렵다. 사람은 한 눈에 스팸을 구별할 수 있지만 컴퓨터를 통한 스팸 디텍션은 쉽지 않다. 이에 본 논문은 웹 스팸을 감지하는 expert를 도와주고자 한다. 스팸을 찾아내는 알고리즘은 매우 어렵기 때문에 사람의 도움이 필요하다. 논문에서 소개하는 알고리즘은 처음에 ‘span status'를 결정할 필요가 있는 작은 seed set page를 정하고 expert가 이 시드 페이지를 살펴봐서 스팸인지 좋은 페이지인지 판단한다. 그리고 좋은 seed page의 connectivity를 이용하여 다른 good page들을 찾아나간다.

Contribution:

1. 웹 스팸의 문제와 스펨 디텍션 알고리즘 소개

2. 디텍션 알고리즘의 효율을 평가할 방법 정의

3. 수동적으로 평가할 seed sets of pages 고르기

4. 믿을만한 페이지인지 TrustRank 알고리즘 적용