인용문헌을 이용한 검색 성능 향상에 관한 실험적 연구

An Experimental Study on the Improvement of Retrieval Performance Using Citation Information

국민상, 정영미 (연세대학교 대학원 문헌정보학과)
Min-Sang Kook, Young-Mee Chung
Department of Library and Information Science, Yonsei University

정보검색시 전문(full-text)의 사용이 늘어남에 따라 다의어나 철자오류와 같은 문제점으로 인해 내부적인 정보원의 사용에 한계를 보이면서 외부적인 정보원, 즉 문헌간의 관계와 같은 링크 또는 인용정보에 대한 관심이 높아지게 되었다. 본 논문에서는 인용링크나 피인용링크, 서지결합링크, 동시인용링크 등과 같은 인용정보와 적합성 피드백 검색을 이용하여 검색 성능을 향상시키는 방안에 대하여 연구하였다.

1. 서론

하이퍼텍스트에서의 하이퍼링크나 논문 등에 등장하는 인용문헌은 텍스트의 내용을 설명해 주는 좋은 지표가 될 수 있다. 특히 웹상의 하이퍼텍스트와 같은 경우는 불분명한 내용이 많고 전통적인 정보검색에서 사용해 왔던 단어 빈도를 효과적으로 사용하는 것이 어렵기 때문에 하이퍼링크와 같은 외부적인 정보원은 매우 유용할 수 있다.

본 논문에서는 인용정보가 벡터공간 모형상에서 검색 성능 향상에 어느 정도 도움이 되는지 실험을 통해 확인해 보고, 검색 성능 향상을 위해 인용정보를 적합성 피드백에서 어떤 방식으로 사용할 것인지에 대해 살펴보도록 한다.

2. 이론적 배경

2.1 인용색인

인용사항을 색인표목으로 사용한 색인을 인용색인이라 한다. 인용색인은 단어와 언어에 대해 독립적이고, 따라서 자연언어에서 발생하는 모호성을 피할 수 있다는 장점이 있다. 또 인용사항 자체가 매우 간결하고 특정 패턴을 따르고 있기 때문에 관리하기가 용이하고, 의미적인 해석도 필요 없으며, 컴퓨터로 처리하기도 용이하다. 그러면서도 인용색인은 문헌간의 의미적인 관계를 설명해 줄 수 있다.

인용문헌을 이용하여 문헌간의 주제적 관계를 설정하는 기법으로는 서지결합 기법과 동시인용 기법이 있는데, 둘 다 문헌의 인용 패턴에 근거하여 문헌간의 관계를 측정하는 기법으로, 내용이 유사한 문헌의 집단화 및 특정한 정보 요구에 대한 문헌검색에 이용되고 있다.

2.2 서지결합 기법과 동시인용 기법

서지결합 기법은 여러 개의 문헌이 공통으로 인용문헌을 하나 이상 가지고 있을 때 이 문헌들은 서로 주제적으로 관련되어 있다는 가정

하에 제시된 것으로, 이 때 문헌들은 서지적으로 결합되어 있다고 말한다(Kessler 1963). 서지적으로 결합된 문헌간의 결합도는 공통으로 인용된 문헌 수로 측정된다. 따라서 결합도가 높을수록 두 문헌의 주제는 유사하다고 본다.

이에 반해 동시인용 기법은 인용문헌을 통한 문헌간의 관계를 서지결합 기법과는 다른 관점에서 파악하는 기법으로 두 편의 인용문헌이 후에 출판된 제 3의 문헌 속에 동시에 인용되었을 때 이 두 편의 문헌들은 서로 주제적으로 관계가 있다고 보는 것이다(Small 1973).

본 논문에서 사용한 기본적인 검색 모형이 벡터공간 모형이기 때문에 인용정보 또한 벡터로 표현되어야 한다. 보통 인용 및 피인용 문헌간의 인용 관계는 <그림 1>과 같은 행렬로 표현된다.

인용문헌 \ 피인용문헌	D2	D3	D5
D1	1	0	0
D3	1	0	1
D4	1	1	1
D5	0	1	0

<그림 1> 인용-피인용 문헌 행렬

이러한 행렬을 A라 할 때, 그 원소 $a_{i,j}=1$은 문헌 D_i가 문헌 D_j를 인용한다는 것을 의미한다. 이 행렬로부터 서지결합정보나 동시인용정보를 얻기 위해서는 행렬 A와 A의 전치행렬 A^T를 곱해주면 된다. 공식 (1)과 (2)에서 $b_{j,k}$와 $c_{j,k}$는 각각 새롭게 생성된 서지결합행렬(BC) 및 동시인용행렬(CC)의 원소를 나타낸다.

$$b_{j,k}=AA^T \quad (1)$$

$$c_{j,k}=A^TA \quad (2)$$

위의 행렬 A로 직접 서지결합 정보와 동시인용 정보를 산출해 보면 <그림 2>와 같다.

$$BC=AA^T=\begin{bmatrix}1&0&0\\1&0&1\\1&1&1\\0&1&0\end{bmatrix}\times\begin{bmatrix}1&1&1&0\\0&0&1&1\\0&1&1&0\end{bmatrix}=\begin{bmatrix}1&1&1&0\\1&2&2&0\\1&2&3&1\\0&0&1&1\end{bmatrix}$$

$$CC=A^TA=\begin{bmatrix}1&1&1&0\\0&0&1&1\\0&1&1&0\end{bmatrix}\times\begin{bmatrix}1&0&0\\1&0&1\\1&1&1\\0&1&0\end{bmatrix}=\begin{bmatrix}3&1&2\\1&2&1\\2&1&2\end{bmatrix}$$

<그림 2> 서지결합 및 동시인용 정보의 산출

위의 서지결합행렬 BC에서 대각 원소는 문헌 D_i가 다른 문헌을 인용한 횟수를, 다른 원소들은 서지결합도를 나타내고, 동시인용행렬 CC에서의 대각 원소는 문헌 D_i가 다른 문헌에 의해 인용된 횟수를, 다른 원소들은 동시인용도를 나타낸다. 이러한 행렬들은 대각 요소를 제외하면 "우상(右上)"의 원소들과 "좌하(左下)"의 원소들이 대칭인 대칭행렬이 된다(Noel 2000).

2.3 인용정보를 이용한 피드백 검색

적합성 피드백 검색은 크게 이용자 피드백 검색과 시스템 피드백 검색으로 나눌 수 있다. 전자는 초기검색 결과 검색된 문헌 중에서 이용자가 직접 적합한 문헌을 선택하면 이를 기반으로 적합한 문헌에 출현한 용어의 가중치는 높여 주고 부적합한 문헌에 출현한 용어의 가중치는 낮추어 줌으로써 질의를 확장하는 방법이고, 후자는 이용자의 개입 없이 초기검색 결과 상위 순위에 검색된 문헌을 분석하여 질의어를 확장하는 방법이다.

이러한 적합성 피드백 검색에서는 적합문헌에 대한 정보를 얻은 후 새로운 질의를 작성할 때 적합문헌에 포함된 모든 용어 또는 가중치가 높은 상위 몇 개의 용어를 포함시키는 것이 일반적이지만, 본 연구에서는 여기에 인용정보를 추가하였다. 즉 1차 검색 후 상위 m개의 문헌 내에서 적합문헌으로 판정된 문헌들로부터 용어 뿐만 아니라 인용링크, 피인용링크, 서지결합링크, 동시인용링크 등과 같은 인용정보

를 함께 추출하여 용어와 인용정보가 결합된 새로운 질의벡터를 생성하는 것이다. 이러한 새로운 질의벡터는 용어 및 인용정보가 포함된 문헌벡터들에 대해 2차 검색에 이용되고, 유사도는 질의벡터와 문헌벡터의 코사인 유사계수로 계산된다.

3. 인용문헌을 이용한 검색 실험

3.1 실험 설계

(1) 실험 문헌집단

인용정보를 포함한 소규모 문헌집단으로 가장 많이 사용되고 있는 것은 CACM 실험집단이다. CACM은 1958~1979년 Communication of the ACM에 게재된 3,204개의 논문으로 구성되어 있고, 컴퓨터 과학에 관한 폭넓은 주제를 다루고 있다. 이 실험집단은 각 문헌에 대한 제목 외에 저자, 날짜, 초록, 분류구조, 키워드, 문헌간의 서지참조, 서지결합, 동시인용 필드 등이 추가되어 있다. 본 연구에서는 이러한 필드 중 제목, 초록, 키워드 필드를 이용해 용어색인 데이터베이스를 만들고, 서지참조, 서지결합, 동시인용 필드를 이용해 인용색인 데이터베이스를 만들어 검색에 사용하였다. 검색 실험에는 총 64개의 자연어 질의 중 적합문헌 정보가 없는 질의와 적합문헌 수가 5개 미만인 질의 등을 제외한 40개의 질의만을 사용하였다.

(2) 검색 모형

색인 대상이 되는 필드들과 질의어는 불용어 제거 후 영문 형태소 분석기인 Porter를 이용하여 색인을 작성하였다. 용어색인 작성시 용어의 가중치($w_{i,j}$)로는 전통적인 TF×IDF(용어빈도×역문헌빈도) 공식에서 TF와 IDF를 각각 최대값으로 정규화된 값을 사용하였다.

$$w_{i,j}=\frac{tf_{i,j}}{\max tf_j}\times\frac{idf_i}{\max idf} \tag{3}$$

$$idf_i=\log\frac{N}{df_i} \tag{4}$$

기본적인 검색 모형으로는 벡터공간 모형을 사용하였고, 유사계수로는 다음과 같은 코사인 계수를 사용하였다.

$$Sim(D_j,q)=\frac{\sum_{i=1}^{t} w_{i,j}\times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2}\times\sqrt{\sum_{j=1}^{t} w_{i,q}^2}} \tag{5}$$

공식 (5)에서 $w_{i,j}$는 문헌 j에서 i번째 용어의 가중치를, 그리고 $w_{i,q}$는 질의 q에서 i번째 용어의 가중치를 나타낸다.

(3) 적합성 피드백

본 연구에서는 시스템 피드백 검색과 인용정보를 함께 이용할 때 어느 정도의 성능 향상을 보이는지를 실험해 보았다. 일반적인 시스템 피드백 검색에서는 공식 (3), (5)와 같은 가중치 공식과 유사계수 공식을 이용하여 1차 검색을 한 후, 상위 m개의 문헌을 적합문헌이라 가정하고 용어색인과 인용색인을 이용하여 2차 검색을 하게 되지만, 본 연구에서는 2차 검색시 인용정보를 추가하기 위해 공식 (6)과 (7)에서와 같이 3쌍의 질의 및 문헌 벡터를 생성하였다. 여기에서 link1과 link2는 인용링크와 피인용링크, 또는 서지결합링크와 동시인용링크 쌍을 나타낸다.

$$Q_{new}=(Q_{new}^{term}, Q_{new}^{link1}, Q_{new}^{link2}) \tag{6}$$

$$D_j=(D_j^{term}, D_j^{link1}, D_j^{link2}) \tag{7}$$

새로운 질의벡터에서 용어 부분은, 질의어 확장만을 사용할 때는 공식 (8)과 같은 Ide Regular 알고리즘에서 $a_1=a_2=1$로, $a_3=0$으로 설정하였고, 인용정보를 이용할 때는 원래의 질의어 Q를 그대로 사용하였다 ($Q_{new}^{term}=Q$).

$$Q_{new}^{term}=a_1Q+a_2\sum_{j=1}^{n_1}R_j-a_3\sum_{j=1}^{n_2}S_j \tag{8}$$

새로운 질의벡터의 인용정보 부분은 인용 및

피인용 링크의 경우 0 또는 1의 값을, 서지결합링크와 동시인용링크의 경우는 각각 서지결합도와 동시인용도를 할당하였다.

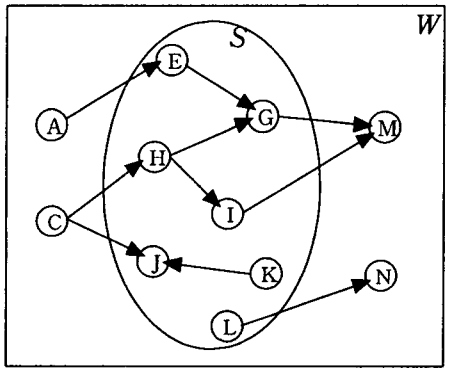
이러한 질의 및 문헌 벡터를 이용하여 재검색을 할 경우 다음과 같은 공식이 적용된다.

$$Sim(Q_{new}, D_j) = \alpha \times Sim(Q_{new}^{term}, D_j^{term}) + \beta \times Sim(Q_{new}^{link1}, D_j^{link1}) + \gamma \times Sim(Q_{new}^{link2}, D_j^{link2}) \quad (9)$$

공식 (9)에서 $\alpha+\beta+\gamma=1$이고, $\beta=\gamma=0$인 경우 인용정보를 고려하지 않은 일반적인 적합성 피드백 검색이 된다.

(4) 인용문헌 필터링

시스템 피드백의 경우 선택된 상위 m개의 문헌 중 적합문헌과 부적합문헌이 섞여있을 수 있고, 그러한 문헌들에 대한 인용 및 피인용 문헌에는 성능 향상에 전혀 도움이 되지 않는 잡음이라고 할 수 있는 인용정보도 포함되어 있을 수 있다. 따라서 본 논문에서는 시스템 피드백 검색을 위해 약간의 필터링 작업을 추가하였다. 즉 검색 결과로 나온 문헌들에 비추어 봤을 때 각 문헌에 연결되어 있는 인용 또는 피인용문헌 중 식별력이 없는 인용 또는 피인용 문헌들은 제거하는 것이다. 그 방법은 다음과 같다.



<그림 3> 검색 결과 S와 인용망

전체 문헌집합을 W, 검색 결과로 나온 문헌집합을 S라 할 때 <그림 3>과 같은 인용망이 있다고 하자. 예를 들어 피인용문헌벡터의 경우 우선 피인용문헌 리스트를 생성한다. 그리고 여기에 S에 있는 모든 문헌을 추가한다. 그 결과로 나온 피인용문헌 리스트는 {E, G, H, I, J, K, L, M, N}이 된다. 이 리스트에서 피인용 횟수가 2번 미만인 문헌은 식별력이 없다고 보고 이 문헌들을 리스트로부터 제거한다. 예를 들어 N은 피인용 횟수가 1번 뿐이므로 리스트로부터 제거한다. 하지만 M의 경우는 G와 I로부터 인용되고 있기 때문에 그대로 남겨 둔다. 마찬가지로 E, H, K, L도 집합 S에서 어느 것으로부터도 인용되고 있지 않기 때문에 리스트로부터 제거한다. 하지만 G, I, J는 집합 S에 있는 문헌으로부터 적어도 한 번은 인용되고 있기 때문에 그대로 남겨 둔다. 이러한 과정이 끝나면 최종적으로 피인용문헌 리스트에는 {G, I, J, M}이 남게 되고 다음과 같은 피인용벡터가 생성되게 된다(Modha and Spangler 2000).

	G	I	J	M
E	1	0	0	0
G	1	0	0	1
H	1	1	0	0
I	0	1	0	1
J	0	0	1	0
K	0	0	1	0
L	0	0	0	0

<그림 4> 필터링 후의 피인용벡터

인용문헌벡터의 경우도 이와 같은 방법으로 생성할 수 있다.

(4) 평가 방법

본 논문에서는 검색 성능 평가를 위해 11-지점 평균 정확률과 R-정확률을 사용하였다. 11-지점 평균 정확률은 정보검색 분야에서 가장 일반적으로 사용되는 평가 방법으로서 검색 모형의 전체적인 성능을 평가하기 위해 사용하였고, R-정확률은 상위 검색된 문헌에 대한 평가를 위해 사용하였다.

3.2 실험 결과

본 연구에서 사용한 시스템 피드백 검색에서

는 1차 검색된 문헌 중 상위 25개의 문헌을 적합문헌으로 사용하였는데, 그 이유는 각각 5개, 10개, 15개, 20개, 25개, 50개의 상위 문헌을 적합문헌으로 하여 검색한 예비 실험 결과 m 값이 20 또는 25일 때 전체적으로 가장 좋은 성능을 보였기 때문이다.

<표 1>은 1차 검색 후, 인용정보 없이 상위 25개의 문헌에 출현하는 용어만을 사용하여 2차 검색한 결과를 나타낸 것이다. 초기검색에 비해 11-지점 평균 정확률은 2.2% 증가했지만, R-정확률은 오히려 1.3% 감소했음을 알 수 있다.

<표 1> 피드백 검색 결과 : 질의어 확장

	11-지점 평균 정확률	향상률	R-정확률	향상률
초기검색	0.334	-	0.319	-
시스템 피드백	0.342	+2.2%	0.315	-1.3%

(1) 인용 및 피인용 링크

<표 2>는 시스템 피드백 검색시 인용 및 피인용 링크를 사용했을 때의 11-지점 평균 정확률과 R-정확률을 나타낸 것이다. α, β, γ 값은 공식 (9)에서와 같이 각각 용어, 인용링크, 피인용링크의 중요도를 나타내는 매개변수에 해당하고, T와 O는 인용 및 피인용 문헌 필터링을 할 때 사용한 매개변수로서 T는 검색된 문헌집단 내에 있는 문헌의 인용 또는 피인용 횟수를, 그리고 O는 검색된 문헌집단 밖에 있는 문헌의 인용 또는 피인용 횟수를 의미한다.

인용 및 피인용 링크를 이용한 시스템 피드백 검색에서는 α 값을 0.9에서부터 0.6까지 0.1씩 감소시켜 가며 실험을 실시하였다. 이 값들 사이에서 검색 성능은 점점 향상되다가 다시 감소되는 경향을 보였고, α 값을 0.6 이하로 낮추면 성능 변화가 거의 없거나 초기검색 성능보다 오히려 더 떨어지는 결과를 보였다.

시스템 피드백 검색에서 인용 및 피인용 링크를 사용할 경우 초기검색에 비해 11-지점 평균 정확률에서는 최대 7.6%(α=0.7; T≥2, O≥3), R-정확률에서는 최대 13.8%(α=0.6; T≥2, O≥4)의 성능 향상을 얻을 수 있음을 알 수 있다. 특히 검색된 문헌집단 내에 있는 문헌의 인용 또는 피인용 횟수가 1 이상이고, 검색된 문헌집단 밖에 있는 문헌의 인용 또는 피인용 횟수가 1 이상인 경우(T≥1, O≥1), 즉 인용 및 피인용 문헌에 필터링을 사용하지 않은 경우보다는 적당한 필터링을 해 주었을 때 현저한 성능 향상이 있음을 보이고 있다.

(2) 서지결합 및 동시인용 링크

<표 3>은 서지결합 및 동시인용 링크를 이용한 시스템 피드백 검색 결과를 나타낸다. 각 표에서 α, β, γ는 공식 (9)에서와 같이 각각 용어, 서지결합링크, 동시인용링크 중요도를 나타내는 매개변수에 해당된다. 서지결합 및 동시인용 링크를 이용한 시스템 피드백 검색에서 서지결합링크나 동시인용링크를 독립적으로 사용할 경우에는 α 값으로 0.8과 0.9를 사용하였고, 둘을 함께 사용할 경우에는 α 값으로 0.7~0.9의 값을 사용하였다. 그 이하의 값을 사용할 경우에는 성능 향상이 거의 없거나 오히려 감소하는 경향을 보였다.

<표 3>에서 11-지점 평균 정확률에서는 초기검색에 비해 서지결합링크를 사용한 경우 0.8%(α=0.9), 동시인용링크를 사용한 경우 2.8%(α=0.9), 둘 모두를 사용한 경우 2.4%(α=0.9)의 성능 향상을 보이고, R-정확률에서는 서지결합링크를 사용한 경우 0.9%(α=0.9), 동시인용링크를 사용한 경우 4.9%(α=0.9), 둘 모두를 사용한 경우 3.9%(α=0.9)의 성능 향상을 보이고 있다.

4. 결론

본 논문에서는 인용링크나 피인용링크, 서지결합링크, 동시인용링크 등과 같은 인용정보를 이용하여 검색 성능을 향상시키는 방안에 대하

<표 2> 인용 및 피인용 링크를 이용한 시스템 피드백 검색 결과

	$\alpha+\beta+\gamma=1$; $\beta=\gamma$	α=0.9	α=0.8	α=0.7	α=0.6
11-지점 평균 정확률	T≥1, O≥1	0.342(+2.4%)	**0.347(+3.8%)**	0.343(+2.7%)	0.334(-0.1%)
	T≥1, O≥2	0.346(+3.5%)	0.356(+6.1%)	**0.358(+6.7%)**	0.349(+4.2%)
	T≥1, O≥3	0.345(+3.2%)	0.354(+5.7%)	**0.359(+7.0%)**	0.347(+3.7%)
	T≥1, O≥4	0.344(+2.9%)	**0.350(+4.5%)**	0.348(+4.0%)	0.342(+2.4%)
	T≥2, O≥2	0.346(+3.5%)	**0.356(+6.2%)**	**0.356(+6.2%)**	0.349(+4.2%)
	T≥2, O≥3	0.346(+3.5%)	0.355(+5.9%)	**0.361(+7.6%)**	0.352(+5.2%)
	T≥2, O≥4	0.346(+3.3%)	0.353(+5.3%)	**0.357(+6.4%)**	0.350(+4.6%)
	T≥3, O≥3	0.344(+2.9%)	0.349(+4.4%)	**0.351(+5.0%)**	0.342(+2.4%)
	T≥3, O≥4	0.343(+2.8%)	0.344(+3.0%)	**0.349(+4.3%)**	0.338(+1.2%)
	T≥4, O≥4	0.344(+3.0%)	0.343(+2.7%)	**0.347(+3.7%)**	0.330(-1.1%)
R-정확률	T≥1, O≥1	0.330(+3.4%)	0.337(+5.4%)	**0.340(+6.1%)**	0.337(+5.4%)
	T≥1, O≥2	0.340(+6.0%)	0.342(+6.7%)	0.351(+9.0%)	**0.351(+9.1%)**
	T≥1, O≥3	0.334(+4.5%)	0.346(+7.7%)	0.344(+7.3%)	**0.347(+8.0%)**
	T≥1, O≥4	0.339(+5.8%)	0.345(+7.7%)	0.350(+8.7%)	**0.353(+9.6%)**
	T≥2, O≥2	0.340(+6.0%)	0.350(+8.9%)	0.352(+9.3%)	**0.363(+12.2%)**
	T≥2, O≥3	0.337(+5.4%)	0.347(+8.2%)	0.359(+11.1%)	**0.361(+11.7%)**
	T≥2, O≥4	0.348(+8.2%)	0.354(+9.9%)	0.364(+12.3%)	**0.370(+13.8%)**
	T≥3, O≥3	0.336(+5.0%)	0.341(+6.4%)	**0.360(+11.4%)**	0.355(+10.2%)
	T≥3, O≥4	0.348(+8.4%)	0.352(+9.2%)	0.360(+11.4%)	**0.362(+11.9%)**
	T≥4, O≥4	0.347(+8.2%)	0.342(+6.7%)	**0.352(+9.4%)**	0.350(+8.9%)

<표 3> 서지결합 및 동시인용 링크를 이용한 시스템 피드백 검색 결과

	서지결합 (γ=0)		동시인용 (β=0)		서지결합 및 동시인용 ($\beta=\gamma$)		
	α=0.9	α=0.8	α=0.9	α=0.8	α=0.9	α=0.8	α=0.7
11-지점 평균 정확률	**0.337 (+0.8%)**	0.331 (-1.0%)	**0.344 (+2.8%)**	0.329 (-1.4%)	**0.342 (+2.4%)**	0.342 (+2.3%)	0.335 (+0.4%)
R-정확률	**0.322 (+0.9%)**	0.310 (-2.8%)	**0.336 (+4.9%)**	0.323 (+1.3%)	**0.332 (+3.9%)**	0.329 (+3.1%)	0.325 (+1.8%)

여 연구하였는데, 시스템 피드백 검색에 인용정보를 이용할 경우 기본 검색 모형이나 질의어 확장을 이용한 시스템 피드백 검색보다 더 나은 성능 향상을 보였다. 하지만 실험에 사용한 실험집단의 규모가 비교적 작고, 실험 문헌집단 내에서 인용정보를 포함하는 문헌 수도 적은 것은 아쉬운 점으로 지적된다.

참 고 문 헌

Kessler, M. M. 1963. "Bibliographic Coupling betwwen scientific papers". *American Documentation*, 14 : 10-25.

Modha, D. S., and W. S. Spangler. 2000. "Clustering hypertext with applications to web searching". *Proceedings of ACM Hypertext Conference.* San Antonio, Texas.

Noel, S. 2000. *Data Mining and visualization of Reference Associations : Higher Order Citation Analysis.* Ph. D. diss., University of Louisiana.

Small, H. 1973. "Co-citation in the scientific literature: A new measure of the relationship between two documents". *Journal of the American Society for Information Science*, 24 : 265-269.