情　報　学　基　礎　55－7
デジタル・ドキュメント　19－7
（１９９９．７．１６）

用語間関係に着目した文書間関係に関する統計的分析と分析支援システムの開発

石田和成*、太田敏澄+
*松江工業高等専門学校情報工学科
ishida@it.matsue-ct.ac.jp
+電気通信大学大学院情報システム学研究科
ohta@is.uec.ac.jp

要旨：本稿では、用語間関係にもとづいた類似度関数を定式化することによって、社会情報学という新たに発展している学術的領域において、中心的な論文や、論文の特徴を明らかにする方法を提案する。中心的な論文を見つけることや、論文の特徴を明らかにすることは、その新たな学術的領域を理解し発展させるのに役立つと考えられる。これまでの研究で、異なる学術的背景の論文から、用語体系を視覚化することは、新しい学術的領域において使用されている用語体系における意味の類似点や相違点を明らかにするのに有効な手段であることを確認した。しかし、意味を明らかにするために視覚化を用いることは、幾分主観的で時間もかかる。そこで、本研究では、類似度関数を用いることによって、類似の用語や、新しい学術的領域で中心的な論文、そして論文の特徴を同定し、できるだけ低い労力で論文間関係を明らかにする。

On a statistical analysis of relation among documents in terms of terminology and a development of system to support the analysis

Kazunari Ishida[+] and Toshizumi Ohta[*]
[+] The Department of Information Engineering, the Matsue National College of Technology
14-4 Nishiikumachou, Matsueshi, Shimane 690-0865, Japan
[*] The Graduate School of Information Systems. the University of Electro-Communications
1-5-1 Choufushi, Choufugaoka, Tokyo 182-8585, Japan

Abstract: In this paper, we propose a method for identifying characteristics of papers and the central papers in the social informatics that is an emerging academic discipline by formulating a similarity function based on terminology. Identifying the most important papers is useful for understanding the new discipline quickly. From our previous research, we are convinced that being able to visualize terminology drawn from papers with different academic backgrounds is an effective method for clarifying similar and dissimilar meanings in the terminology used in the new discipline. However, using visualization to clarify meaning is slightly subjective and time consuming. By employing the similarity function, we were able to identify similar terms, which papers were central to a new discipline we examined, and clarify the relations between papers without excessive effort.

1.はじめに

本稿では、用語間関係に着目した文書間関係に関する統計的分析と分析支援システムの開発について報告する。分析対象としては、社会情報学に関する基礎的な論文[1][2]を用いる。これは、研究者が専門分野に関して、一貫した概念体系を持っており、その論文において用いられている用語間関係は、そのような概念体系の表現系の１つと考えられるためである。そのため、本論文では、文書間で、用語間関係に関するいくつかの統計量を定義し、それにもとづき、論文間の類似度を定義する。用語間関係は、用語の用いられているコンテクストを表現するものと考えられるため、多義的な意味を持つ用語は、

各研究者がもつ概念体系をつなぐ重要な要素であると考えられる。このような用語間関係を用いた分析は、情報組織化のための基礎的な方法となると考えられるため、社会情報学のような、新しい学問分野における概念体系の整備や、学問体系の発展に伴い成長する事典の構築に役立つと考えられる。

2.情報組織化に関する従来の研究の問題点

本節では、情報組織化に対する、本研究のアプローチと、従来の研究のアプローチとの違いを述べる。具体的には、情報を組織化するためには、従来の研究が情報検索におけるノイズの原因として扱っていた多義的な意味をもつ用語が重要であることを述べる。

2.1.情報検索の基礎的方法

情報検索(Information Retrieval)を行うためには、文書における情報を組織化するために、キーワードの抽出や、ベクトル空間における文書の配置を行うことが基本的な方法となっている(Salton,1989)。これらの方法を用いた情報検索において、用語の持つ意味の多義性を削減することは、情報検索の研究における重要な課題のうちの 1 つである。用語の持つ意味の多義性によって、情報検索の利用者は、キーワードによる情報検索結果の爆発に苦しむこととなる。

この多義的な意味を持つ用語を除去するために、Term frequency and inverted document frequency (tf・.idf) 法がよく用いられる。この方法は、文書の特徴を良く表す用語を文書中から抽出する方法である。この用語抽出の後、各用語を軸としたベクトル空間を構成し、その空間に文書を位置付けることによって情報組織化を行うことができる。また、クラスタリングなどの手法を用いて文書を分類することができる。

Chen & Lynch (1992) は、生物学、あるいは医学という良く整備された情報に関して、概念空間を構築するために、tf・idf 法やクラスタリングなどの方法を用いて、生物学や医学に関する概念のネットワークを構築した。Schunze (1992, 1998) は、クラスターに用語の意味を割り当てるために、クラスタ内の少数の事例を参考として、クラスタと語義の対応づけを行っている。Yarowsky (1992) は、用語間の共起関係における矛盾の無いように、単一の意味を持った用語を用いてクラスタリングを行っている。

2.2.概念間の関連づけとしての多義的用語

しかし、多義的な意味を持つ用語は用語間関係の観点から、概念間の関係を明らかにするときの鍵となるものであると考えられる。例えば、社会情報学おいて、「社会」や「情報」は基礎的な用語となっている。社会情報学に関心のある人々は、これらの用語がどのように用いられているのかが知りたいであろうと考えられる。たとえば、社会科学系の研究者が「情報」をどのように捉えているのかということを、情報工学系の研究者は知りたいであろうし、その逆の場合も考えられる。もしも、これらの用語を、tf.・idf 法で処理すると、「社会」や「情報」は全ての論文に含まれている基本的な用語であるため、索引用語としては選ばれないこととなる。これに対して、本研究では、これらの用語が、各論文において各著者がどのように用いているのかを明らかにすることが、新しい学問領域の基盤整備のために重要となると考えている。

2.3.多義的用語と文脈

用語、あるいは用語間関係を使用する文脈は、用語の意味を決定するように思われる。例えば、「組織」という用語が、「社会」や「経営」などの用語と共起する場合、この「組織」という用語の意味は、「人間によって構成される組織」という意味であると考えられる。あるいは、「心臓」、「肝臓」などの用語と共起する場合には、「細胞によって構成される組織」という意味であると考えられる。次節では、用語が使用される文脈にもとづいて、文書における用語体系を分析し、論文の特徴や、論文間の関係を明らか

にする方法を提案する。

3.論文間関係の分析

本節では、社会情報学に関する論文を組織化するための本研究のアプローチを説明する。物事を組織化するとき、類似性、非類似性という尺度は基礎的な要素の１つであると思われる。そのため、類似性の定義は、情報組織化が成功するか否かを決める鍵となる。

3.1.用語体系間の類似性の定義

社会情報学に関する論文を組織化するために、本研究では論文の著者の持っている用語体系に着目する。なぜなら、用語体系の多様性は、社会情報学という新たに発展しつつある学問分野において必然的に生じるためである。

文章は、ある意味で相互に関連のある用語の集合であると考えられる。そのため、本研究では、文章における用語の共起関係を用いることによって、用語間関係を定義する。

もしも、文書１と文書２において、用語 A と用語 B がそれぞれ共起関係にある場合、文書１と文書２での、用語 A の用法の類似性は、高いものと考えられる。ここで、このテキストの関係を、「共通用語関係(Common Term Relation、CTR)」と呼ぶこととする（図１）。

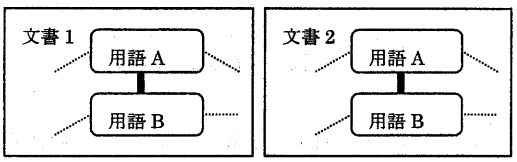

図 1.　共通用語関係（CTR）

しかし、文書間の用語の共起関係にはもう一つの側面がある。例えば、文書１と文書２が、用語 A と用語 B をそれぞれ含んでいる状況を考えよう。もし、用語 A と用語 B は文書１では共起しているが、文書２では共起していない場合、これらの文書間での用語 A の用法の類似度は低いものと考えられる。これは、用語の用法の非類似性と捉えることができる。この文書間関係を、「排他的用語関係（Exclusive Term Relation, ETR)」と呼ぶこととする（図２）。

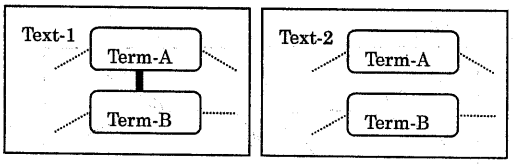

図 2.　排他的用語関係（ETR）

3.2.類似性関数の定式化

論文間の類似性を調べるために、論文間での用語の共有度や、3.1 で定義した CTR や ETR を用いて、次の 3 種類の統計量を、2 つの論文間で抽出した。それらは、(1)用語重複度(Term Overlap Ratio、TOR)、(2)用語共起関係度（Common Term Relation Ratio, CTRR)、そして、(3)用語排他的関係度（Exclusive Term Relation Ratio, ETRR）である。

TOR は 2 つの文書間の、用語の共有度を示している。TOR は次のように定式化した。TOR=|CT|/(|TS1|+|TS2|-|CT|)。CT は 2 つの文書間での共通用語の集合である。TS1 と TS2 は、それぞれ文書１と文書２における全ての用語の集合である。|CT|は CT における要素の数を表している。図３は CT、TS1,そして TS2 の関係を示している。

CTRR は、2 つの文書の間の用語の用法の類似度を表している。CTRR は次のように定式化した。CTRR = |CTR|(|TRS1| + |TRS2|-|CTR|)。CTR は、図１に示した、文書間の共通用語関係の集合である。TRS1 と TRS2 はそれぞれの文書での用語間関係の集合である。図４は CTR,TRS1,そして、TRS2 の関係を示している。

ETRR は 2 つの文書間での用語の用法における非類似性を示している。ETRR は次のように定式化する。ETRR=|TER1／TRS1|＋|TER2／TRS2|。ここで TER1 は 2 つの文書間での排他的用語関係の集合である。この用語関係は図２で示した。TER2 は TER1 と同じような方法

で求めることができる。TRS1 と TRS2 は CTRR の定式化において説明した。図 4 は TER1,TSR1,TER2,そして TRS2 との関係を示している。

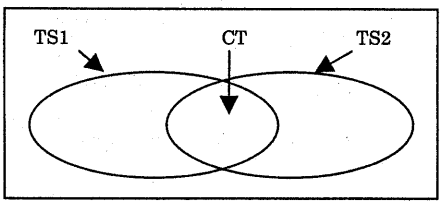

図 3. CT,TS1,TS2 の間の関係

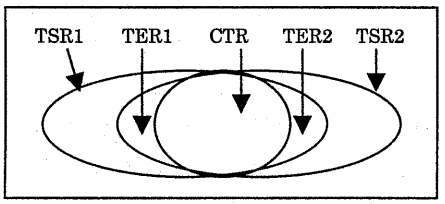

図 4.CTR,TSR1,TSR2,TER1,TER2 の間の関係

ここで定式化した、TOR、CTRR、そして ETRR にもとづいて、類似度関数 S = TOR + a * CTRR − b * ETRR を定義する。ここで、a と b は値として 0 か 1 を取るパラメータである。このパラメータを用いて、類似度計算において、CTR と ETR を考慮するかどうかを指定できる。

3.3.T-Analyzer と D-Analyzer

本研究では、用語間関係と文書間関係を抽出するために、Term Analyzer (T-Analyzer) と Document Analyzer (D-Analyzer)とを開発した。T-Analyzer は TSi と TSRi をそれぞれの文書から抽出する。D-Analyzer は、それぞれの 2 つの文書間での組み合わせにおいて、CT, CTR, TSR1, TSR2 を抽出し、類似度を関数 S にもとづき計算する。次節では、この 2 つ Analyzer を用いることによって、類似度について TOR のみを考慮する場合、TOR と CTRR を考慮する場合、そして TOR,CTRR,ETRR を考慮する場合それぞれについて、用語間関係や文書間関係を分析する。

4.分析結果

社会情報学に関する 8 つの論文に対して、分析を行った。分析は次の 3 通りとした、1 つは、共通用語の割合（TOR)、2 つ目の分析は、共通用語間関係の割合（CTRR）と TOR、3 つ目の分析は、排他的用語間関係の割合（ETRR）と TOR、CTRR を用いた分析である。このように、用語間関係を考慮することによって生じる類似度の変化を調べる。

分析に先立ち、各論文の統計量を示すと以下のようになる。この図は、用語の出現頻度 tf、用語間関係の出現頻度 trf、そして、用語間関係の出現数と可能性のある全ての用語間関係の数との比 trf/ptr、という 3 つの観点から、文書の特徴を示したものである。これら 3 つの統計量を 1 つのグラフで表現するため、データの正規化を行っている。

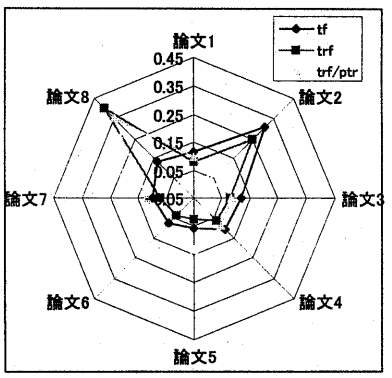

図 5．各論文の統計的特徴量

この図から、論文 2 と論文 8 が、他の論文の特徴と著しく異なることがわかる。論文 2 は、他の論文と比べて、用いている用語数が非常に多く、用語間関係も非常に多いということである。しかし、用語間関係比は、他の論文と比べて最も低くなっている。これは、この論文が、用語間関係の観点から「疎」であるということを示している。

これに対して、論文 8 は、用語数に関して、他の論文と大差はない。しかし、用語間関係が著しく多くなっている。これに伴い、用語間関

係比も、他の論文と比べて著しく高くなっている。これは、この論文が、用語間関係の観点から「密」であることを示している。また、論文４も、他の論文と比べて用語間関係が密であることがわかる。

4.1.分析 1(TOR)

TOR による類似度評価では、論文１と論文３の類似度が 1.0 と最も高かった。つづいて、論文２と論文 8 の類似度が 0.641 と高かった。これらの論文の組は、同一研究者によるものであるためと考えられる。

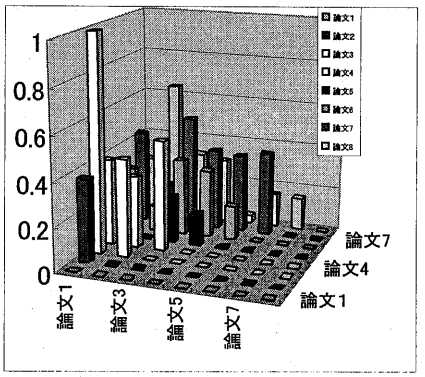

図 6.類似度マトリックス

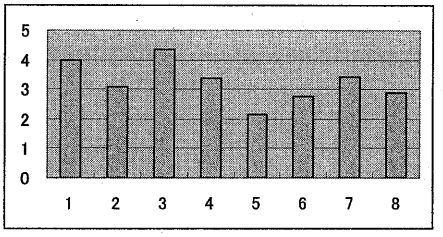

図 7.各文書毎の総類似度

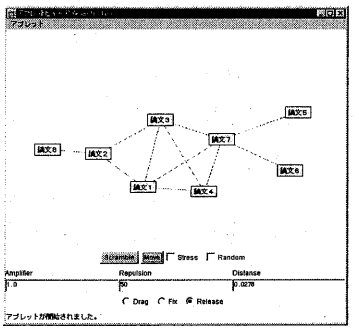

図 8.分析１による論文間類似関係

また、著者は異なるが、論文３と論文４、論文３と論文７の間の類似度も、0.509、0.520 と比較的高い。

逆に、類似度が低い論文間関係には、論文２と論文５が 0.00、論文２と論文６が 0.114 などがあげられる。

4.2.分析２（TOR&CTRR）

TOR&CTRR による類似度評価でも、分析１と同様に、論文１と論文３の類似度が 1.0 と最も高かった。

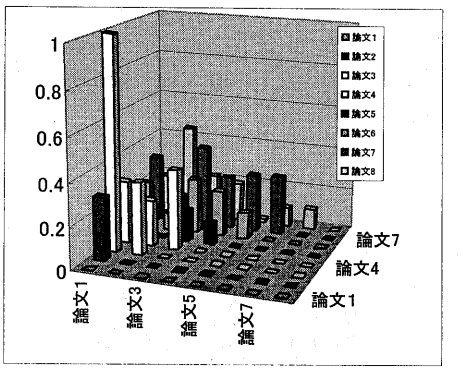

図 9.類似度マトリックス

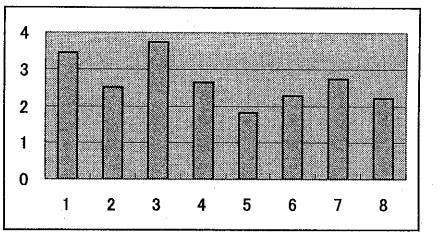

図 10.各文書毎の総類似度

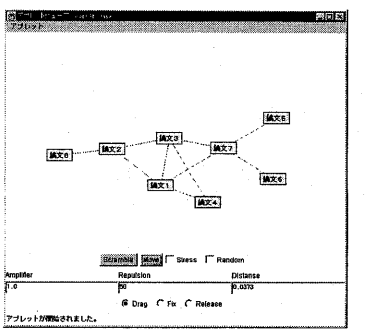

図 11.分析２による論文間類似関係

また、論文２と論文８の類似度は 0.442 と、前回と比べ下がったものの、他の論文間関係と比べ相対的に高かった。さらに、論文３と論文４、論文３と論文７の間の類似度も、0.379、0.385 絶対値は下がったものの、相対的な順位は分析１の結果と同様に高い。

類似度が低い論文間関係も、依然として、論文２と論文５が 0.00、論文２と論文６が 0.08 となっている。

4.3.分析３（TOR,CTRR、&ETRR）

TOR,CTRR,&ETRR による類似度評価でも、分析１、２と同様に、論文１と論文３の類似度が 1.0 と最も高かった。また、類似度が低い論文間関係は、論文２と論文５が 0.002、論文２と論文６が０となっている。

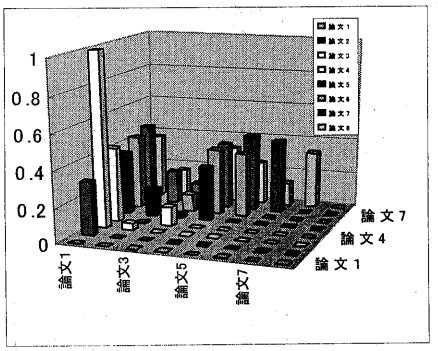

図 12.類似度マトリックス

しかし、論文２と論文８の類似度は 0.166 と、大きく降下した。さらに、論文３と論文４、論文３と論文７の間の類似度も、0.106、0.127 と降下し、相対的な順位も低下している。

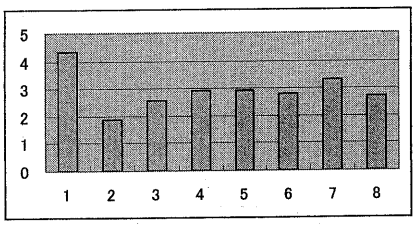

図 13.各文書毎の総類似度

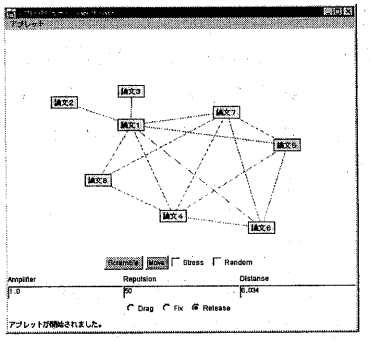

図 14.分析３による論文間類似関係

5.考察

本節では、４節で得られた分析結果の考察を行う。まず 5.1 では、各分析１,2,３の間の関係について考察する。5.2 から 5.6 では、それぞれ特徴のある論文を取り上げ考察を行う。

5.1.各分析の間の関係について

分析１と分析２を比較すると、CTRR を導入することによって、TOR のみの場合の結果の性質を大きく変えるものではないことが分かる。類似度の差が変化しているように見えるのは、文書１、３が同一著者に記述されているため、その類似度が増大したことによるものである。

しかし、分析３と他の分析との結果を比較すると、質的に大きな違いが見られる。例えば、論文２と論文８の類似度は 0.166 と、他の分析結果と比較して、大きく降下している。

5.2.同じ単一著者による論文について

２つの論文１と論文３が共通著者１人で記述されている場合、常にその類似度が高いことが、図６、９，１２から分かる。図１５は、これらの論文間の用語間関係を、T-Visualizer を用いて、社会、情報、システムという用語を論文のキーワードとして、それらの用語に直接結びついている用語を視覚化したものである。

この図では、左側に論文１の、社会、情報、システム、という用語を、右側に論文３の、社

会、情報、システム、という用語をそれぞれ配置している。これらの用語の間で、意味、社会情報、社会情報システム、情報システム、環境、は 2 つの論文間で、3つのキーワードに対して同じ共起関係を持っている。このように、この論文間では、CTR の割合が高いことが視覚的に把握できる。

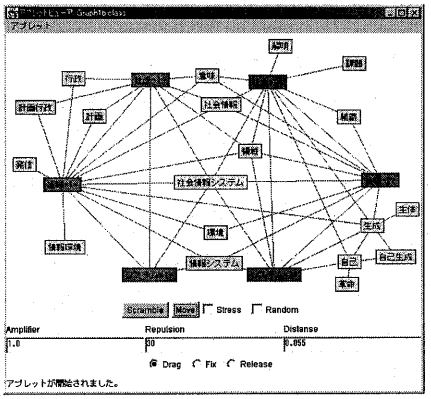

図 15. 論文１と論文３の用語の共有関係

5.3.複数著者による論文に対する ETRR の影響について

論文２と論文８は、この 2 つの論文は同一著者が記述したものであるため、分析１、分析２では、類似度の総体順位は 2 位と高かった。しかし、論文８は、その著者を含む 3 人の共同著作となっているため、3 種類の用語の用法が用いられている可能性がある。

１つの論文における複数の用語体系の使用は、図５で示した論文の特徴量 trf/ptr に影響を与えている。論文２のこの値は、一番低くなっており、用語間関係が疎となっている。

そのため、他の論文との間で、ETRR の値が大きくなってしまったものと考えられる。このように、ETRR を類似度の評価に用いることによって、複数著者の論文を検出できる可能性がある。

図１６は、「社会」「情報」という用語を中心的用語として、それらと関連ある用語を図示することによって、2 つの論文が、どのような用語を共有しているかを示している。この図からわかるように、論文間で主要用語を直接結ぶ用語の割合が低いことがわかる。

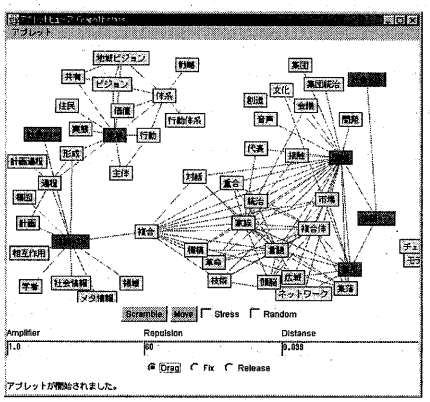

図 16. 論文２と論文８の用語の共有関係

5.4.解説論文に対する ETRR の影響について

また、図１０と図１３を比較するとわかるように、ETRR を類似度関数に導入することにより、論文３の総類似度が相対的に低下している。この原因の 1 つとして考えられるのは、論文３の内容に関係があると考えられる。論文３は、社会情報全般に関する解説的な論文であり、詳細な理論展開やシステムの説明などを行う文書ではない。そのため、図５に示すように、trf/ptr が低くなっている。その結果、他の論文との間で、ETRR の値が大きくなってしまったものと考えられる。このように、ETRR を類似度の評価に用いることによって、解説論文を検出できる可能性がある。

5.5.論文のサイズと論文間類似度におよぼす ETRR の影響

文字の分量が相対的に少ない論文は、必然的に用語の数|TS|と用語間関係の数|TSR|が小さくなるため、分析１と分析２では、他の論文との相対的な類似度が低くなる傾向がある。この傾向は、例えば、論文５に関する結果において現れている。分析１（図７）と分析２（図１０）では、論文５の総類似度は最も低くなっている。しかし、ETRR を導入することによって、

分析３（図１３)に示すように、相対的な類似度は上昇している。これは、3.1 における ETR の定義から、この比が 1 に近ければ近いほど、ETRR の値は０に近づくため、trf/ptr の値が高い論文５は、ETRR による類似度への負の影響を受けにくくなるためである。そのため、類似度の計算において、論文サイズの影響を押さえることができると考えられる。

5.6.論文８に対する ETRR 値の影響

しかし、論文８は trf/ptr 値が非常に高いにもかかわらず、ETRR によって相対類似度はあがっておらず、逆に低下している。この原因は、文書の中に、大きな表があり、この表にある用語の間に全ての用語間関係があるものとして用語間関係を抽出してしまい、本来共起しない用語を、共起するものとしてしまったためと考えられるこれにより、他の論文とも、ETR の増大を招いていると考えられる。そのため、図５に示すように、trf の値と、trf/ptr の値が、他の論文と比べて非常に高い値となっている。

5.7.まとめ

本節の考察から、用語間関係 CTR と ETR を考慮すると、文書間の類似度によって、(1)複数著者による文献が判別できる可能性がある（論文２)。(2)解説論文とそうでない論文を判別できる可能性がある（論文３)。(3)論文の分量が少なくても、論文間関係がある程度適切に類似度を計算できる可能性がある（論文５)、と考えられる。

6.結論

本稿では、用語間関係に着目した文書間関係に関する統計的分析と分析支援システムの開発について報告した。分析対象としては、社会情報学に関する基礎的な論文を用いた。このような分析は、社会情報学のような、新しい学問分野における概念体系の整備や、学問体系の発展に伴い成長する事典の構築に役立つと考えられる。また、用語間関係は、用語の用いられているコンテクストを表現するものと考えられるため、研究者の持つ概念体系を、客観的に捉えるための基本的な要素となると考えられる。

参考文献

[1] 社会情報システム学コロキウム編、「社会情報システム学･序説」、富士通ブックス、1996.

[2] 小林宏一、浜田純一、太田敏澄 他著、「社会情報学のダイナミズム」、富士通ブックス、1997.

[3] Ishida, K. & T. Ohta, "On a Visualization System of Terminology Relations in Digital, Documents -- Toward Developing an Encyclopedia of Social Informatics --," The 5th symposium of Information Systems for Society, pp. 19 - 24、1999.

[4] Ohta, Toshizumi and Kazunari Ishida, "Toward a Development of an Encyclopedia of Social Informatics," *Proc. of the 43nd Annual Conference of the International Society for the Systems Sciences* (CD-ROM), 1999, forthcoming.

[5] Chen, H. and K. J. Lynch, "Automatic construction of network of concepts characterizing document database, IEEE Transaction on Systems, Man and Cybernetics, 22(5):885-902, September/October 1992.

[6] Salton, G., "Automatic Text Processing," Addison-Wesley, Reading, MA, 1989

[7] Schutze, H, "Dimensions of Meaning," in Proc. of Supercomputing, pp. 787-796, 1992.

[8] Schutze, H, "Word Sense Discrimination," Computational Linguistics, Vol. 24, No. 1, pp. 97-123, 1998.

[9] Yarowsky, D., "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," in Proc. Of COLING, pp.454-460, 1992.