デジタルドキュメントにおける共起データを用いた検索ターム連想支援について

徳田 圭世*、間瀬 久雄**、辻 洋**

日立西部ソフトウェア(株)*
(株)日立製作所 システム開発研究所**

文書情報の検索に全文検索が使われることが多くなってきた。検索にはユーザによる適切な検索タームの入力が不可欠であるが、タームの選択によってはノイズが検索されたり漏れが発生するため、結果をみて、絞り込みのタームを加えたり、別のタームを選択する必要がある。我々は、このターム選択（連想）を支援するために文書から抽出した共起データを格納するシソーラス管理システムを開発し、それがユーザにどのように貢献するか、どのように提示するのがよいかを明らかにするため、被験者を立てて連想実験を行った。タームの連想という観点からはシソーラスにより約４８％連想タームが増えること、検索インタフェースという観点からはターム提示にクラスタリングが必要であることを確認した。

Search term association support with co-occurrent data from the digital documents of search target

Tamayo Tokuda*, Hisao Mase**, Hiroshi Tsuji**
Hitachi Seibu, Ltd.* Hitachi, Ltd.**

People have become to search the data and information from a large amount of electronic documents. For effective search, it is needed for user to input appropriate search term, but depending on the selected term, the user sometimes face noise data or lack of necessity data. We consider that related terms to the search term are useful to support imagining the adding term, so we have developed thesaurus management system, which stores the extracted co-occurrent data from the document set. We evaluated the association experiment with the subjects, for the purpose of clearing how the co-occurrent terms contribute to user and how indication is better. We have assured that the number of associated terms were 48% increased in the point of term association, and that clustering of indication data was needed in point of retrieval interface.

1．はじめに

ネットワーク時代を迎え，情報が複雑化／大規模化する中、電子化された情報を取得するために検索エンジンが使われるようになってきた。当初の検索エンジンは、目的情報を特徴づけるタームを入力し、ＡＮＤやＯＲの論理演算を駆使して所望の結果を得られるように設計されていた[1]。

しかし、一般のユーザは論理演算などを指定することが困難であり、検索結果にノイズが多く含まれたり、逆に多くの漏れを生じることが多い。我々は、こうしたノイズや漏れに対してユーザが適切なタームを入力することを支援する研究を行っている。

研究の一つのアプローチとして、シソーラスを用いる方法がある。従来は、上位語、下位語をシソーラス設計者が作成してきた[2][3]が、最近では検索対象文書から、同一文書に現れやすい言葉の組を共起語としてシソーラスに格納しようという動きもある[4][5]。共起語をシソーラスに格納するメリットとして、以下があげられる。

(1)あるタームを入力したときに大量の結果が得られた場合、共起語で絞り込めば間違いなく絞りこめるし、絞り込み結果の件数を事前に掌握することも可能である

(2)共起語を順次たどることにより、連想の支援をすることが可能である

しかしながら、これはあくまで期待であり、共起語が実際ユーザにとってどのように有効であるか、提示方法をどのように工夫すればユーザにとって有効であるかはあまり論じられてこなかった。

本研究では共起語データを用いた連想支援について 20 人の被験者をたてて、実験を行った。WWWホームページ約 45000 件から抽出した共起語 6 万組を用い、被験者を立てて実験した。以下では共起語が連想に与える影響、連想したタームの種別の分析、被験者の意見に基づく検索インタフェースの設計ガイドラインについて考察を与える。

2． 共起語による検索ターム連想支援

2.1 検索ターム連想支援イメージ

絞り込み用の情報をユーザに提示する方法として、我々は入力タームの共起語を用いるアプローチを採用している。検索対象ドメインの文書群から抽出した共起語を用いると、一定数以上のヒット文書が保証でき、また確実な絞り込みが可能である。ユーザが思い付いた検索タームで絞り込み検索を行う場合、追記するタームの表現が少し違うだけでヒット数が０件になったり、逆に多すぎたりという場合がある。検索に対し適確な表現の言葉を提示することにより、このようなケースは減少し、また追記するタームの連想が困難な場合に支援が可能である。

我々は、必要に応じた情報をユーザに提示するため、共起語も含め特定の単語と強く関係する関係語の情報を一括管理するシソーラス管理システム Theater を用いて、検索ターム連想の支援を目指す[6]。

2.2 シソーラス管理システム Theater の概要

本システムは、以下の機能を持つ汎用ツールである。

(1)既存シソーラスコンテンツの登録

(2)文書群から共起語情報を抽出・登録

(3)(1)(2)で登録したデータの一括管理

より検索対象に近い分野の既存シソーラスを登録する、検索対象文書群から共起語を抽出し登録するなど、目的に応じた運用を行う。

本稿で報告する実験では、既存の EDR 辞書データと、約 45000 件の WWW ホームページから抽出した共起語情報とを Theater に登録し

た。連想支援ツールとして、登録データ中から入力タームの周辺語、階層語、共起語を各々リスト表示するものを用いた。Theater による支援イメージを図 2.1 に示す。

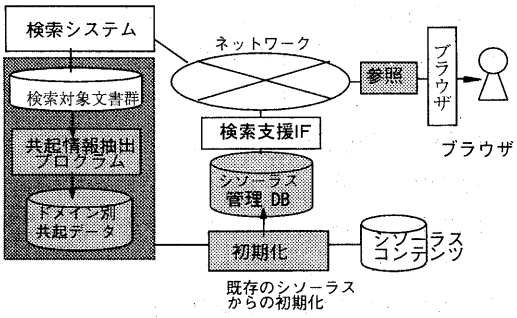

図 2.1 Theater による検索支援イメージ

3．検索タームの連想実験

3.1 実験目的

本実験はタームの連想支援における共起語の有効性検証を目的としている。具体的には、主に以下に示す項目を考察する。

(1)提示データとしてどのようなタームが有効か（i.e. データを提示したとき被験者はどのようなタームを選択するか）

(2)どのような検索インタフェースが必要か

実験で用いる支援データの提示ツールを図 3.1 に示す。

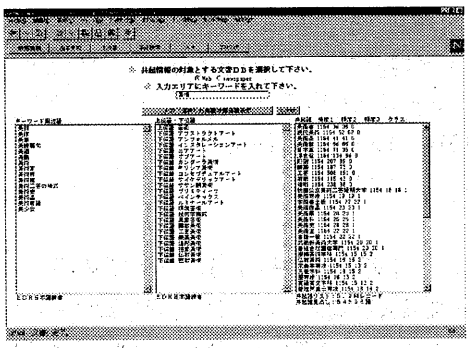

図 3.1 表示ツールの画面例

3.2 実験方法

3.2.1 基本タスク

実験条件は異なるが、被験者に与える基本的なタスクを以下に示す。

［課題］与えられたキーワード（以下、課題ターム）に対し連想した言葉（以下、連想ターム）をワークシートに記入する。

［条件］タスク遂行の基本は 3 分間。連想が尽きたら途中終了もＯＫ。続行も可能だが 5 分過ぎれば強制終了する。

［注意点］連想タームが課題タームからかけ離れないよう指導する。

実験で用いたワークシートを図 3.2 に示す。計算機上での記入は入力作業の負荷が実験結果に影響すると考え、紙ベースの実験とした。

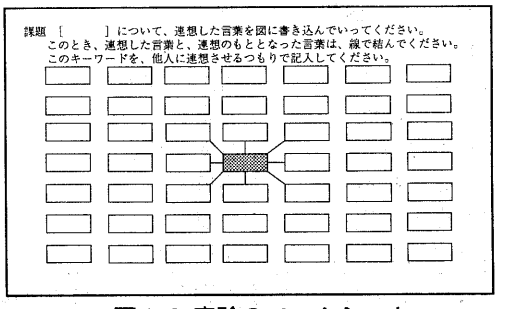

図 3.2 実験のワークシート

事前の予備実験で、3 分間の課題遂行で途中終了の被験者が複数いたこと、「これ以上出てこない」旨の発言が複数聞かれたことから、3 分間を妥当と判断した。

連想タームが課題タームから離れすぎないために、「記入した連想タームを見て他人がもとの課題タームを連想できるように」という条件をつけた。

3.2.2 課題タームの選定

被験者の課題タームに対する得意・不得意による連想ターム数への影響を除くために、Hole-in-One、NTT ディレクトリなどWWWディレクトリサービスのカテゴリを参考に 14 分野を作成し、関心度が高い／普通／低い分野を各 2 ～ 3 個選択してもらった。課題タームは、関心度が高い／普通／低い分野から各 1

題、計3題を選定する。そのため被験者毎に課題タームが異なる。

課題タームは、予め14分野毎に挙げておいた課題ターム候補からランダムに選択する。分野とそれに対応する課題ターム候補一覧を図3.3に示す。

＝分野＝	＝課題ターム候補＝
趣味・生活	：趣味、家庭
娯楽・ｴﾝﾀｰﾃｲﾒﾝﾄ	：娯楽、チケット
市場・ｺﾝﾋﾟｭｰﾀ	：サービス、通信
芸術・アート	：美術、デザイン
旅行	：旅行
ｽﾎﾟｰﾂ・ｱｳﾄﾄﾞｱ	：ｽﾎﾟｰﾂ、ｱｳﾄﾄﾞｱ
ﾏｽﾒﾃﾞｨｱ・情報	：メディア
教育	：学校、図書館
健康・医療	：医療、健康
科学・学問	：科学
自然環境	：農業、宇宙
行政・政治	：国家
歴史・文化	：伝統
社会制度一般	：法律

図3.3 分野と対応する課題ターム候補一覧

3.2.3 実施方法

実験A、実験Bの2種類を実施する。実施条件を表3.1に示す。

表3.1 実施条件

###	実験A	実験B
課題数	3題	3題
被験者の群分け	なし	実験群 コントロール群
被験者数	20名	10名／群

実験Bでは、実験の目的とする要因を実験条件に付加した「実験群」と、その要因を排除した「コントロール群」の2つの被験者群を用意した[7]。群分けは、(1)年齢層(2)ブラウザ利用への慣れ(3)課題遂行力の面を考慮して均等に振り分ける。

以下に、各実験の実施手順を示す。

[実験Aの手順]

(1)課題タームに対して、ツールを使わずに3分間基本タスクを行う。

(2)次に、同じ課題タームに対して、ツールを使って3分間基本タスクを行う。ツールを使わずに行ったワークシートに追記する。

(3)実験後、ツール利用に関する感想についてインタビューする。

(4)実験中は観察記録を取る。

[実験Bの手順]

(1)課題タームに対して、ツールを使わずに3分間基本タスクを行う。

(2)10日±1日間の実験間隔をあける。

(3)同じ課題タームに対して、実験群は最初からツールを使って6分間、コントロール群はツールを使わずに3分間基本タスクを行う。別のワークシートに新たに記入する。

(4)実験後、ツール利用に関する感想についてインタビューする。

(5)実験中は観察記録を取る。

なお、実験群の課題遂行時間は、ツール使用の時間を見積もって6分間とした。

3.3 評価指標

両実験に共通して以下の評価指標を用いる。

(1)被験者別平均個数

(2)全体平均個数

(3)関心度分野別平均個数

(4)課題タームと連想タームとの意味関係

さらに実験Bでは、

(5)連想タームの再現率

について調べる。

4. 連想実験の結果分析および考察

被験者の個人差は、(1)課題に対するハードルの高さ(2)発想力(3)課題レベル(4)課題に対する得意不得意、などと考えられる。被験者別にデータを分析する場合以外は、個人差を考慮に入れた考察が必要になる。

4.1 実験Aの結果分析および考察

(1)被験者別平均個数

連想タームの被験者別平均個数図 4.1 に示す。図 4.1 の「基本データ」は最初の連想タームを示し、「Theater 使用」がツール利用時も含めた連想タームを示す。

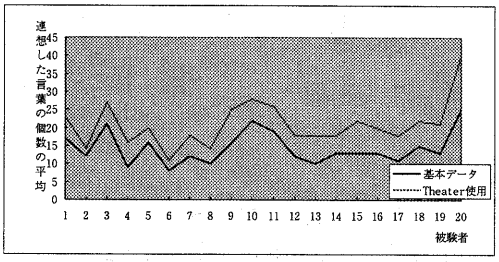

図 4.1 被験者別平均個数の比較

[結果] 全ての被験者において関連データの提示による連想ターム増加が見られた。

(2)全体平均個数

連想タームの平均個数を表 4.1 に示す。

表 4.1 全体平均個数

1回目の連想ターム平均数	14.5個
連想ターム増加数の平均	6.6個
平均増加率	48%

[結果] 平均増加率 48%という数値を得た。すなわち、提示データが連想タームの想起に有効であった。

(3)関心度分野別平均個数

関心度分野別の分析結果を図 4.2 に示す。ただし、特異な結果を出した被験者1名のデータは除く。

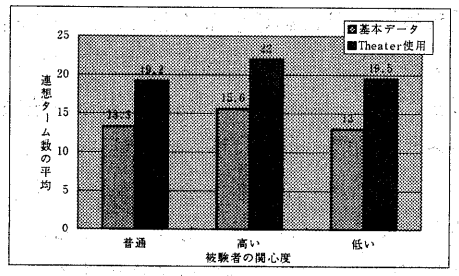

図 4.2 関心度別平均 (特異データ除く)

[結果] 最初の連想ターム数および増加した連想ターム数に関心度別差異は認められない。

[考察] 連想ターム数は関心度に左右されていない。しかし、実験で用いた課題タームは一般的なタームであり、専門的なタームを用いれば関心度別差異が見られる可能性もある。

(4)課題タームと連想タームとの意味関係

被験者 20 名に各3題、計 60 題の全データについて、課題タームと連想ターム間の「is-a 関係」「has-a 関係」「同義語関係」「複合語関係」を調べた。その個数および比率を表 4.2 に示す。なお本稿では、課題タームを部分文字列として含むタームを「複合語」とする。

表 4.2 連想タームと課題タームの関係

###	最初の連想ﾀｰﾑ中の総数	増加ﾀｰﾑ中の総数	増加比率
Is-a	132	47	35.6%
Has-a	8	3	37.5%
同義語	19	2	10.5%
複合語	119	92	77.3%

[結果] 平均増加率 48%と比べて複合語の増加比率が高い。連想支援情報として複合語が適していることが伺える。

[考察] データ提示により連想タームに質的変化がなければ、各関係語の増加比率も 48%付近と予測できる。逆に増加比率から提示データの影響を測ることが可能である。

実験の結果、複合語の増加比率が高く、被験者が提示データ中の複合語を連想の参考にしたことが伺える。実際、Theater が格納する共起語は複合語を扱っており、これを裏付けている。

4.2 実験Bの結果分析および考察

(1)被験者別平均個数

被験者別の連想ターム平均個数を、実験群は図 4.3 に、コントロール群は図 4.4 に示す。

[結果] コントロール群から、実験が 2 回目であることによる影響は無視できるという結果を得た。実験群の6分間の実験で連想タ-

ムが出尽くす範囲が広がった。

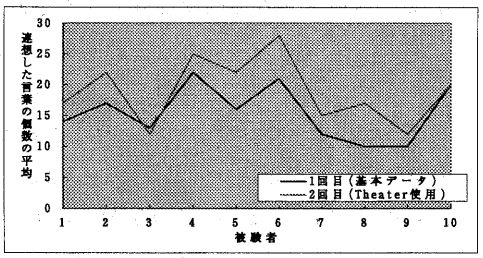

図 4.3 被験者別平均（実験群）

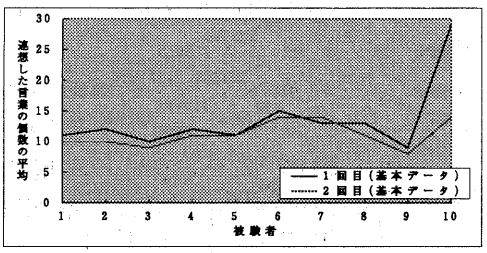

図 4.4 被験者別平均（コントロール群）

［特異値に関する考察］

コントロール群には特異値が1つ存在する。該当する被験者はインタビューで「2回目は初回より質を上げようとした」と言及していること、観察記録によれば、初回は下位語を列挙しているという所感を得たが2回目はそうでなかったことから、特異値とするのが妥当と考える。これ以外のデータからは、「2回目の実験」であることによる影響はみられない。したがって実験群の分析では、実験が2回目であるという要因を無視できると考える。

(2)全体平均個数

連想タームの平均個数を表 4.3 に示す。

表 4.3 全体平均個数

＃＃＃	実験群
1回目連想ターム平均数	15.6個
2回目連想ターム平均数	19.2個
平均増加率	23%

［結果］実験Aと比較して、平均増加率の低下が見られる。

［考察］平均増加率の低下要因として、2回目の実験では連想タームと提示される支援情報とが混在し情報が錯綜した可能性がある。

(3)関心度分野別平均個数

関心度分野別平均個数を図 4.5 に示す。

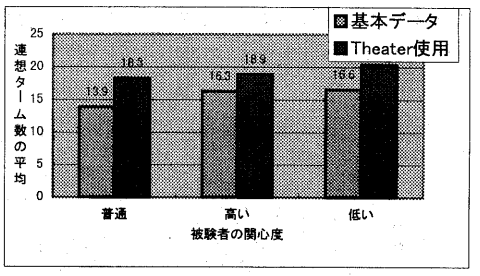

図 4.5 関心度別平均（実験群）

［結果］実験Aと同様、有意差はみられない。

(4)課題タームと連想タームとの意味関係

2回の実験に共通する連想タームは、認知レベルが高いタームと考えられる。それ以外のタームについて課題タームと連想タームの意味関係を調べ、比較の結果を表 4.4 に示す。

表 4.4 連想タームと課題タームとの関係

＃＃＃	本群		コントロール群	
	差分ﾃﾞｰﾀ中の総数（初回）	差分ﾃﾞｰﾀ中の総数（2回目）	差分ﾃﾞｰﾀ中の総数（初回）	差分ﾃﾞｰﾀ中の総数（2回目）
Is-a	41	42	30	16
Has-a	3	5	2	1
同義語	0	1	0	0
複合語	15	46	27	36

［結果］（コントロール群の値を基準として）複合語、Is-a関係のタームが増加した。

［考察］結果に影響する要因は、ツール利用による影響または時間的要因の2つである。観察記録によればツール利用の少ない被験者が多く、前者の影響は小さいと見られる。

ツール利用の有効性を見出せない場合、被験者は頭の中だけで連想しようとする。6分間の実験時間は頭の中だけで連想するには長く、無理に捻り出そうとするため、従来の連想領域からはずれたタームが発想されて質変

化が生じたものと考える。

(5)連想タームの再現率

初回の連想ターム数に対し、2回とも連想したターム数の比率を「再現率」と定義し、分析した結果を表 4.5 に示す。ただし、特異データは除いた。

表 4.5 再現率

###	実験群	ｺﾝﾄﾛｰﾙ群
全課題平均	40.6%	43.0%
最大再現率	71.4%	75.0%
最小再現率	12.5%	12.5%

[結果] 連想ターム中の平均 40%が課題タームに対して認知度の強いタームである。

[考察] 結果からは両群間に再現率の差は見られないが、実験群では連想タームが出尽くしていることを考えると、コントロール群と比較してやや再現率が低い点が気にかかる。被験者が本来連想する範囲のタームを書き切る前にツールの情報が視覚に入るため、連想が途中で遮断されてしまう可能性がある。

4.3 分析結果および被験者意見に基づく検索インタフェースの検討

4.3.1 分析結果の考察

(1)提示タイミング

最初から支援情報があると、ユーザ本来の連想を妨げる恐れがある。表 4.1 および 4.3 によると、実験Aの平均増加率 48%に対し実験Bは 23%である。被験者の意見でも、実験Aでは有用性を認める意見が聞かれたのに対し、実験Bでは比較的少なかった。

被験者の連想タームを広げる観点からは、実験Aのタイミング即ち連想が尽きた頃に支援情報を提示する方が良いと考える。

(2) 同義語支援の必要性

表 4.2 によると、ユーザが自然に連想する場合、Is-a 関係のタームや複合語が多く連想されるのに対し、同義語は連想タームとしては出現しづらいようである。しかし、同義語は検索においてその役割が重要視されており、同義語支援についても検討が必要である。

4.3.2 被験者意見の考察

(1) 提示方法について

[主な被験者意見]

・提示タームが多く、選ぶのが大変。
・関連語の提示が思考範囲を限定する。
・共起語が整理されておらず見づらい。

[考察]

共起データには有用な情報だけでなくノイズも含まれているため、何を見せるかというデータの選択と共に、提示データをクラスタリングして見やすくする工夫が必要である。

また、提示データが思考範囲を限定するという意見には2つの側面があると考える。検索の連想支援であれば、検索ドメインに応じて連想の方向づけや枠組みを与えることは有効と考えるが、思考を妨げる方向に働いてはならない。

(2)ツール利用について

[主な被験者意見]

・使用義務意識が働く。ツールがない方がいい連想ができる。提示されるタームに頼る。
・馴染みの薄い課題タームに対し役立った。連想タームを思い出す手間が省けた。

[考察]

使用に対する義務意識は、ユーザの思考がツールにより妨げられていることを示している。ユーザ作業を妨害しないインタフェース設計を検討すべきである。

また、馴染みの薄いテーマについて効果が見込まれる。逆に、思い出す手間や入力を省く観点からも有用性を目指す。

4.3.3 インタフェースの考察

(1) 共起語のクラスタリング

共起語のクラスタリングの必要性、および複合語の有効性を示す実験結果を得た。例え

ば検索ターム「保険」に対する複合語「生命保険」「保険会社」などは絞り込み検索を有効に支援すると考えるが、「検索タームを部分文字列に含むターム」はクラスタリングの1つキーになると考える。

(2) ユーザ目的の把握

本来のユーザ目的を妨げないインタフェース設計が必要であることを実験で確認した。例えばユーザの検索目的の度合いを把握し、それに応じたインタフェースを提供することが有効に作用する可能性がある。検索目的の度合いを把握するための一方法として、例えば最初から2つ以上の検索タームを入力するユーザは明確な目的を持っており、それに対して1つだけタームを入力する場合、目的が比較的あいまいという見方ができる[8]。

例えば後者の場合、(2)で示した複合語情報はその有効性が期待されるが、前者の場合は検索方向とずれる可能性がある。被験者の入力パターンによって提示タームを変える工夫ができると考える。

5．まとめ

我々はシソーラス管理システム Theater を用い、被験者を立てて共起データによる検索タームの連想支援実験を行った。その結果、情報提示による連想タームの個数が約 48%増加すること、課題タームを部分文字列として含む複合語が連想タームの発想面で有効であることを確認した。検索インタフェースについては、共起語のクラスタリングが必要であること、連想が尽きたタイミングで支援データを提示する方が良いことが分かった。

情報の連想関係を評価する従来研究には、情報に含まれるタームを関連付けた「連想構造」でその情報を構造化する手法に関する研究がある。上記研究では、連想構造に基づいたターム検索が有効であるという結果を得ている。しかし、連想構造を利用したシステム側の視点での評価であり、ユーザを主体とした発想支援という視点での研究はなされていない。

我々はユーザにとって連想支援という観点でどのような情報が有効であるかの研究が必要と考え、ユーザが連想できるタームの範囲を広げる連想実験を、共起語を用いて行った点に特徴がある。

謝辞

本研究を行う上で、共起情報の利用方法、その抽出プログラムの提供などでご指導いただいた（株）日立製作所基礎研究所主任研究員丹羽芳樹博士に感謝します。

6．参考文献

[1]DOORS、pp.18-20、朝日新聞社、1996-3

[2]JICST 科学技術用語シソーラス 1993 年版、日本科学技術情報センター、1993

[3]日経シソーラス、日本経済新聞社、1994

[4]間瀬久雄他：WWW ホームページからの共起語自動抽出実験、情報処理学会第 55 回全国大会論文集(3)pp.72～73、1997-9

[5]井上孝史他：追加検索語候補提示に関する一検討、情報処理学会第 55 回全国大会論文集(3)pp.76、1997-9

[6]徳田圭世他：情報検索支援のためのシソーラス管理システムの提案、情報処理学会第53回全国大会(3)pp.163-164、1996-9

[7]谷津進他：実験計画法 pp.60、培風館

[8]P.イングベルセン著、藤原鎮男監訳：情報検索研究 pp.240、トッパン

[9]前田晴美他：弱い情報構造に関する評価実験、人工知能学会全国大会(第 11 回)論文集 pp.344-347、1997-6