社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (1)
2006－ＮＬ－175 (1)
2006／9／12

書籍の目次と索引を利用した専門用語ネットワークの構築

石塚 隆男
亜細亜大学経営学部

ある特定の専門用語の定義的な意味だけでなく、どういう状況や文脈のもとで用いられているかを知りたい場合にそれが索引語になっている専門書が役に立つ。そこで、本研究では、書籍の目次と索引を活用し、専門用語ネットワークを構築し、可視化することを目的とする。今回、統計学と自然言語処理関係の書籍を対象に処理を行い、いくつかの知見が得られたので報告する。索引語と目次のマッチングを行うことにより索引語である専門用語を包含する上位概念を階層的に得ることができ、同時に共起的に用いられている他の専門用語を把握することができる。専門用語×書籍マトリクスから、書籍の特性が得られ、書籍マップや特徴語の抽出が行えることを示した。目次と索引は、いわゆる文章形式で書かれたものでないため、日本語形態素解析はほとんど必要なく、一般語は除外されているので質の高い情報として有用であろう。

A Visualization of Technical Terms Network using Table of Contents and Indexes in Books

Takao Ishizuka
Asia University

We propose a new method for constructing a network of technical terms. Although the Internet contributes to the propagation of vast electronic texts, it doesn't provide structured knowledge possible to take into our brains easily. Our method utilizes a table of contents and an index list of each book. Contents have hierarchical structures and enable us to know the contexts and superordinate concepts of index terms. We collect plural technical literatures per one domain and construct terms vs. contents matrices by matching contents and indexes. Several knowledge visualizations are possible by basic operations of relational database. One of the merits of utilization of contents and indexes is the high quality of information without bothering with general words comparing with morphologic analysis of raw sentences.

1．はじめに

今日、デジタル化は当然のこととなっているが、複数の媒体を用いずにパソコン画面からほしい情報にアクセスできるようになったにすぎず、ネット上から得られるものは“半形式知”である。情報検索によりヒットした文書は得られるが、それらを読み、自らの頭の中で構造化する作業は人間が行わなければならない。たとえば、自分にとって未知の専門用語の意味を知りたい場合、専門書や事典等を調べることにより解決する場合もあるが、むしろ私たちが知りたいのはその専門用語がどのような状況あるいは文脈の中で用いられているかを知りたいことの方が多い。同様に、私たちはある概念の意味を正確に記憶しているよりもど

ういう状況でその概念を用いるかを知識としてもっていることの方が多い。たとえば、“インストール”の定義的な意味を正確に説明できなくても、パソコンで新しいソフトを用いる際に行う準備作業のことを言うことは知っており、それで十分である。概念は一連の対象を一言で表現することを可能にするが、いくらその辞書的な意味を知っていても使いこなせばければ意味がなく、意味以上に“文脈的用例”が重要であるといえよう。

本研究では、こうした問題意識から、書籍についている目次と索引を利用することにより専門用語のネットワークを構築することを目的とする。なお、本研究では、専門書の索引に掲載されている索引語を専門用語と定義する。索引語がすべて専門用語とはいえないが、ある分野における重要な専門用語は必ず索引に収載されているはずである。

今回、統計学関係と音声・自然言語処理・情報検索の領域について専門用語ネットワークの構築・可視化を試みた結果、いくつかの知見が得られたので報告する。

2．目次と索引の利用について

本の目次や索引の意義や効用について今更説明する必要はないだろうが、専門用語ネットワークを構築するのに専門書の目次と索引を利用した研究はほとんど見当たらない。

目次により、著者がどのような構成でその本を著そうとしたのか、著者の枠組みあるいは書籍のタイトルが示す学問領域の全体像を知ることができる。多くの本の目次は章・節レベルまであり、目次自体にも専門用語が含まれるが、大きな上位概念により記述されているのが普通である。一方、索引は具体的な専門用語について記述箇所を見つけるためのツールである。１冊の本をデータベースにたとえれば、目次と索引はマスターデータセットであり、本文はバリアブルデータセットとみることができる。

本研究では、マスターデータセットに相当する目次と索引を電子化し、ページ番号によるマッチングにより専門用語のネットワークを構築するものであり、以下にその方法を述べる。

3．方法

1）準備作業

ある特定の学問分野に関する書籍文献を数冊収集し、専門用語の基礎データとして用いる。

専門分野における事典に関してはCD-ROM版等の電子書籍が存在するが、多くの専門書は紙媒体である。そのため、イメージ・スキャナーとOCRソフトを用い、目次と索引をテキストファイルに変換する。

日本語OCRソフトは、精度が向上したとはいえ、漢字の“二”とカタカナの“ニ”の区別がつかない点や“ル”を“ル”(カタカナの“ノ”＋“レ”)に誤変換するなど人間が目視により修正作業を行わなければそのままでは使えない。

目次は、以下の項目からなるCSV形式に変換した。

シーケンス番号，章・節番号，見出し，ページ開始，ページ終わり

なお、ページ開始と終わりは索引語と突合するために必要な情報であるが、書籍には章・節の終わりページの記載がないので前後のページ番号から手入力を行った。通常、章は改ページされて始まるが、節や項は１ページ内に前節の終わりと次節の始めが共起するのが普通である。索引にはページ番号しか記載されていないため、どちらの節に属する専門用語かはひとつひとつ人間が判断しなければならないが、簡単のため、節の終わりページには次節の開始ページの１ページ前のページ番号を入力した。さらに、章等の上位の目次の開始ページと終わりページは、包含する節や項を含むように入力した。これにより、対象とする書籍の本文はすべて“漏れなく、かつ、重複することなく”、章・節の単位に区分することができ、しかも、専門用語を属する項→節→章→書籍全体の各レベルで把握することが可能となる。

また、索引は段組みで印刷されており、“～を見よ”といった参照情報も掲載されているが、今回は、

索引語，書籍タイトル，ページ番号（複数）

のCSV形式のファイルに変換した。なお、参照ページが“○－○”のように連続したページの形式で示されているものについては、同じ節、項の

中での複数回出現のため今回は最初のページのみ入力した。なお、索引は、読者への便宜を図って書籍ごとに多少個性が見られ、索引語の多寡に影響する。語の転置等により索引に複数回登場する索引語については、重複するものは削除した。

以上の準備作業を、18冊の専門書を対象に行い、以下の作業を容易にするため、各書籍の目次、索引をマージしたファイルを作成した。なお、索引語には同じ用語でありながら日本語表記のわずかな違いや索引語の日本語に（ ）付きで英語表現を付加したものなどが多数存在し、文字列的に別の単語として認識されるのでExcelの一括置換機能によりできる限り表現の統一を行ったが、完全ではない。

2）各目次に上位目次の付加

目次は、書籍タイトル→章→節→項の階層構造をしている。ある階層の目次内容の上位の目次がわかることにより概念間の包含関係を知ることができる。専門用語から、上位概念を知るために属する節、項の目次内容に上位の目次を自動的に付加するプログラムを作成した。

3）専門用語と目次とのマッチング

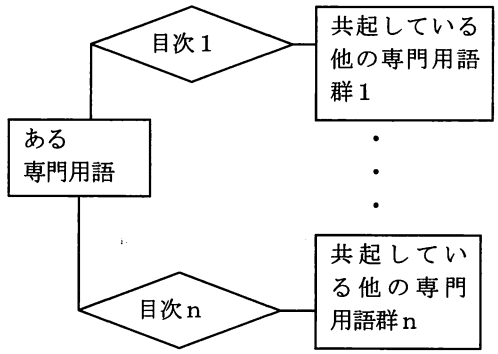
各目次の開始ページと終わりページの情報から索引語が参照するページ番号が該当する目次を探し、各索引語と2）で付加した上位目次までを付加した目次のマッチングファイルを作成した。これにより、複数の参照ページがある索引語が、さまざまな文脈あるいは状況下で用いられていることを容易に知ることができる。

4）専門用語の共起語のネットワークの構築

3）の作業により、各索引語が属する章・節・項が明らかとなったので共起している他の索引語を容易に知ることができる。本研究では、書籍本文の文章解析は行っていないので、概念間の関係まではわからないが、図1のERダイヤグラムに示すように少なくとも共起語がある項、節、章に登場する関連する概念であることはわかる。

従来、“概念マップ”といわれるものは共起度の高い概念を線でつないだにすぎず、どういう文脈により線で関係づけられているのかは利用者が判断するしかなかったが、本研究の方法によれば概念間の関係を目次により集約することができる。

図1．専門用語のネットワークのＥＲ図



また、複数の共起語群の実体をマージし、ひとつの実体にまとめることもできよう。

5）専門用語×書籍マトリクスの作成

ある分野に関する複数の書籍の索引語をマージし、書籍別の出現頻度を集計することにより専門用語×書籍マトリクスを作成することができる。これにより、個々の書籍の特性とともに書籍間の関係を定量化・可視化することができる。作成した主な図表は以下のとおりである。

・専門用語×書籍マトリクスの図解
・索引語数による書籍マップの作成

6）各書籍の特徴ある専門用語の抽出

各書籍の他書、類書にはない特徴を探るために、専門用語×書籍マトリクスの各出現頻度をTF*IDF法により修正を行い、書籍ごとにTF*IDF値の合計（ΣTF*IDF）を算出した。なお、ここでいう特徴とは当該書籍がカバーしている領域の中で他書には記述がほとんどないものを指す。ある書籍の内容が別の書籍の部分集合で各索引語の出現頻度が1の場合には、特に重点的にある分野について書かれた書物とは言えず、特徴がないことになる。特徴語を抽出するのに、書籍の下位グループを構成する目次情報を活用することが考えられるが、高頻度の索引語は複数の章・節で参照されているため、目次情報を集約しにくく、今後の課題である。

一般的に索引語数が多い書籍は、他の書籍には出現していない索引語を多数含み、それらのIDF

図２．専門用語×書籍マトリクスからの書籍別特徴語の抽出方法

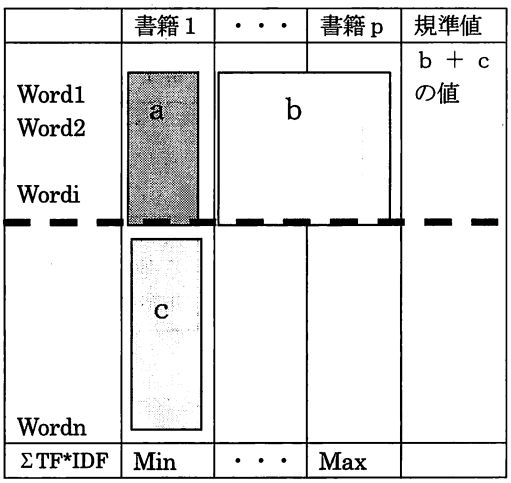
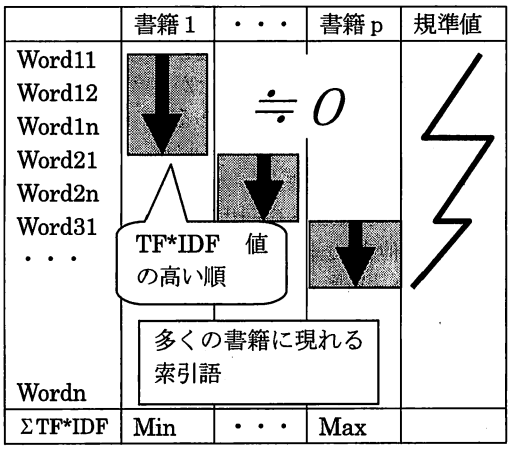


図３．各書籍の特徴語の抽出イメージ



値は高くなるため、必然的にΣTF*IDF 値が高くなる。一方、ある書籍の索引が別の書籍の索引の部分集合でしかも入門的な内容あれば、索引語数が少なく、しかも対応する IDF が小さいのでΣTF*IDF 値は小さくなる。したがって、マトリクスの列項目の書籍集合の中からΣTF*IDF 値の小さい順に書籍を選択し、TF*IDF 値の高い専門用語を以下の方法により抽出する。

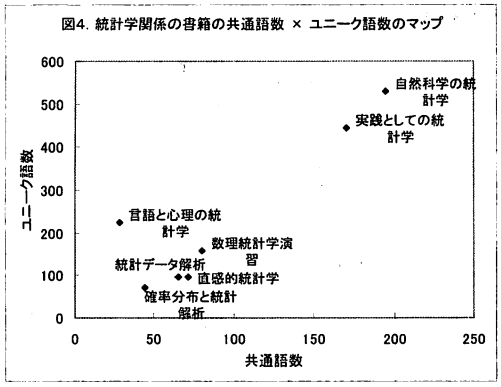
図２は、専門用語×書籍マトリクスであり、書籍はΣTF*IDF 値の小さい順に並んでいるものとする。書籍１の特徴語を抽出するアルゴリズムについて以下に説明する。

マトリクスの各専門用語の行を書籍１の項目の降順にソートする。TF*IDF の性質から上位にランクされた語の書籍２～ｐの出現頻度の大半は０もしくは小さな値である。それらの用語のどこまでを書籍１の特徴語とするかについて本研究では、Word1～Wordi までを特徴語としたとき、マトリクスのｂ領域とｃ領域のTF*IDF値の合計を規準とした。

ｂ領域は、書籍２～ｐでの出現頻度が小さいことを示し、小さければ小さいほどよいが、ｃ領域が大きいのは特徴がないことを意味する。
特徴語を増やすことによりｂの値は単調に増加し、ｃは単調減少する。そこで、ｂ＋ｃが最小値となる単語までを書籍１の特徴とすることが考えられる。以下、書籍２～書籍ｐを対象に同様の作業を行っていくことにより各書籍の他書にはない特徴語を抽出することができる。図３は、上述の作業を逐次的に全書籍について行ったイメージを示したものである。「≒０」と書かれた領域のほとんどの TF*IDF 値は０であり、他書にはない各書籍の特徴語が抽出できていることを示している。基準値の列のノコギリ状のグラフは b+c 規準の値が変化する様を示している。書籍１～p まで作業を行うと最後に多くの書籍に現れる索引語が残る。

４．結果

今回は、巻末に示す統計学関係の書籍７冊、音声・自然言語処理・情報検索関係の書籍 12 冊を選択し、これらの目次と索引をテキストファイルに変換し、分析を行った。うち１冊は両方の分野にまたがるものである。統計学関係の書籍群は、数理統計学の専門書を中心に出版年が古いものから最近のものまでで構成されている。自然言語処理の分野は関連分野や基礎技術が多岐にわたっており、比較的最近の出版物が多い。ここでは音声や情報検索関係の書籍を数冊加え、群を構成した。



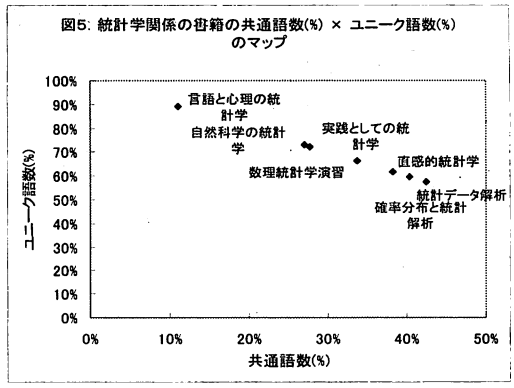


表１．専門用語に出現した目次＋上位目次を付加した結果の例

索引語	頁	目次→上位目次→★書籍タイトル
n-gram	107	内的な評価方法 | テキスト自動要約システムの性能評価 | ★テキスト自動要約
n-gram	127	言語モデル | ★リアルタイム音声認識
n-gram	16	文章の特徴抽出 | 文章の統計分析とは | ★言語と心理の統計学
N-gram	122	言語処理技術の利用 | ★情報検索と言語処理
n－gram model	198	知識の獲得峰畢曲 | コーパスベースの技術 | 自然言語処理の基礎技術 | ★自然言語処
n－gram model	270	統計に基づく翻訳システム | 機械翻訳 | 自然言語処理の応用技術 | ★自然言語処理
N-gramエントロピー	18	言語のエントロピー | 音声言語の確率・統計モデル | ★音声言語処理
N-gramクラス・モデル	35	N-gramモデル | 音声言語の確率・統計モデル | ★音声言語処理
N-gramクラス・モデル	110	単語のクラスタリング | 自然言語処理への応用 | ★音声言語処理
N-gramモデル	27	N-gramモデル | 音声言語の確率・統計モデル | ★音声言語処理
N-gramモデル	90	連続音声認識システムの例 | 音声認識への応用 | ★音声言語処理

表２．目次によって関連づけられた専門用語と共起概念の出力例

n-gram	★テキスト自動要約	7.1.1	内容に関する評価	18	| 11点平均精度 | BLEU | F値 | LCS | n-gram | pseudo utilityに基づく評価 | ROUGE | unigram | utilityに基づくオフライン評価 | スキップバイグラム | 基準要約 | 機械再現率 | 最長共通部分列 | 自動評価 | 精度 | 内容に評価
n-gram	★リアルタイム音声認識	5.2	n-gram言語モデル	8	| n-gram | n-gram言語モデル | カットオフ | コーパス | グラム | バイグラム | ユニグラム | 最尤推定量
n-gram	★言語と心理の統計学	I－2	文章の特徴抽出	8	| 4分位数 | n-gram | TR | 機能語 | 識別語 | 助詞の中央値 | 読点の打ち方
n－gram m	★自然言語処理	6.1	知識の獲得峰畢曲	13	| Kullback－Leibler情報量 | n－gram model | クラス(単語の) | 共起知識の抽出 | 深層格フレーム獲得 | 情報量 | 対訳コーパス | 単語クラスの抽出 | 単語の意味抽出 | 定型表現 | 表層格フレーム獲得 | 翻訳知識の名詞間の類似度の測定

図４は、統計学関係の各書籍について、他書にはないユニーク語を索引語としている数と他のいずれかの書籍に索引語がある語数（＝共通語）をマップで示したものである。図５は、図４の共通語とユニーク語の割合をマップで示したものである。これらの図からユニーク語の内容はわからないが、他書との位置関係をつかむことができる。

たとえば、図４に示すように『自然科学の統計学』は他書に比べ、索引語数が多く、多くの書籍を包含し、他書にはない索引語数でも群を抜いている。図５では、『言語と心理の統計学』が他の統計学にはない索引語の割合が高いことがわかる。

表１は、専門用語に出現した目次＋上位目次を付加したファイルの一部である。N-gram が基本的な技術として、自然言語処理はもとより情報検索、テキスト自動要約等の分野で用いられている

表３．統計学関係の書籍から特徴語を抽出した結果

書籍名	確率分布と統計解析	統計データ解析	直感的統計学	数理統計学演習	言語と心理の統計学	実践としての統計学	自然科学の統計学
ΣTFIDF	101	199	261	226	298	1091	1227
TFIDF値が上位の特徴語	ｷｭﾑﾗﾝﾄ	累積カイ２乗	回帰線	R.A.Fisher	確率的ｺﾝﾌﾟﾚｷｼﾃｨ	データ	繰り返し測定
	ﾌｫﾝﾉｲﾏﾝ	対比較	管理図	回帰関数	SVM	ｐ値	１元配置
	ホテリング	Kruscal-Wallis検定	Ｚ値	２次形式	ｻﾎﾟｰﾄﾍﾞｸﾄﾙﾏｼﾝ	個体差	最尤推定
	許容限界	交互作用	ばらつき	Cauchy 分布	CUSUM	統制	水準
	ガンマ分布	Dunnetの多重比較	季節変動	ｋ変量正規分布	ESC	無作為	漸近分布

ことがわかる。また、索引語の表現が統一されていないこともよくわかる（“N グラム”や“ｎグラム”の表現も用いられている）。

表２は、専門用語の共起語を複数文献から抽出した例である。N-gram について多様な関連語が存在することがわかる。

表３に本研究で述べたアルゴリズムにより統計学関係の各書籍から特徴語の抽出した結果を示す。各書籍の特徴がよく出た結果が得られた。なお、今回は TF*IDF 値の高い上位数語を抽出したが、TF*IDF 値の分布状況を見て判断することが考えられる。

５．考察並びに今後の課題

本研究では、書籍の目次と索引から専門用語間の関係を可視化し、併せて各書籍の特徴を抽出する作業を行った。各索引語の TF*IDF 値を要素とする書籍ベクトル間の類似度をもとに書籍マップの作成を試みたが、専門用語×書籍マトリクスの情報をうまく集約することができなかった。代替的に、本研究では手作業で因子分析を行うのに近い原始的な方法で各書籍の特徴語の抽出を行った。

目次や索引は、一般語がほとんど含まれず、形態素解析が不要であり、データとしての質は高いことから、より多くの書籍から網羅的かつ横断的に活用されるべきと考える。

今回は、索引語の表現の統一が不十分であったこともあり、出現頻度がばらけていることは否定できない。今後、概念間の類似性の判定や書籍間の距離についても可視化を試みたい。

対象とした文献リスト

＊統計学関係

宇野利雄(1955)『数理統計学演習』共立出版

金明哲他(2003)『言語と心理の統計学』岩波書店

佐伯胖他編(2000)『実践としての統計学』東京大学出版会

竹内啓(1975)『確率分布と統計解析』日本規格協会

東京大学教養学部統計学教室編(1992)『自然科学の統計学』東京大学出版会

広津千尋(1983)『統計的データ解析』日本規格協会

吉田耕作(2006)『直感的統計学』日経ＢＰ社

＊音声・自然言語処理・情報検索関係

安藤彰男(2004)『リアルタイム音声認識』電情通学会

奥村学他(2005)『テキスト自動要約』オーム社

鹿野克亘他編著(2001)『音声認識システム』オーム社

岸田和明(1998)『情報検索の理論と技術』勁草書房

北研二他(1996)『音声言語処理』森北出版

北研二(1999)『確率的言語モデル』東京大学出版会

北研二他(2002)『情報検索アルゴリズム』共立出版

田中穂積監修(1999)『自然言語処理』電情通学会

徳永健伸(1999)『情報検索と言語処理』東京大学出版会

長尾真(2004)『言語情報処理』岩波書店

中川聖一(1988)『確率モデルによる音声認識』電情通学会

溝口理一郎(2005)『オントロジー工学』オーム社