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

2005－CH－65 (4)
2005／1／28

異本解析アルゴリズム（SynopticPatch）による計算共観表の作成

三宅真紀、赤間啓之、馬越庸恭*、中川正宣

mmiyake@dp.hum.titech.ac.jp

東京工業大学大学院社会理工学研究科
*東京工業大学学術国際情報センター

概要：本研究では、類似している 2 つの文章から共通部分を含む並行箇所を自動的に切り分け抽出する異本解析アルゴリズムを提案し、その具体的な適用方法として、新約聖書学の「共観福音書」から並行箇所の最適範囲を自動的に出力させた「計算共観表」の生成法について説明する。

The Computed Synoptic Tables by the Algorithm of SynopticPatch

Maki Miyake, Hiroyuki Akama,
Masanori Nakagawa, Nobuyasu Makoshi*

mmiyake@dp.hum.titech.ac.jp

Department of Human System Science, Tokyo Institute of Technology
* Global Scientific Information Center, Tokyo Institute of Technology

Abstract：In this paper, we propose an algorithm partitioning off the texts, called SynopticPatch, which allows us to calculate at every step of the frame extension of the "windowing" method the correlation coefficient between the word frequency vectors generated from each corresponding window instance. For our goal of the scientific examination of the synoptic problem in the New Testament studies, we propose a new statistical method of generating the segmentation criteria of the synoptic Gospels, a sort of "TextTiling" methodology enabling a computed synoptic table (CST) with an objective segmentation based on objective criteria.

1. はじめに

本研究では、単語の一致度に重点を置いた計算共観表を作成することを目的として、類似している 2 つの文書から共通箇所を自動的に切り分け抽出する異本アルゴリズム（SynopticPatch）を提案する。そして、新約聖書の共観福音書から並行箇所を切り出す際の SynopticPatch の具体的な適用方法を示す。すなわち、伝統型共観表を参照しながら共通部分を抽出して並行箇所を設定し、計算共観表を作成する手順について説明する。さらに、この計算共観表の特徴、また伝統共観表の改善点について述べる。

2. 背景

2.1 共観表

新約聖書の文学類型の一つ、福音書について述べる。この文学類型は、キリスト教会において新しく作り出されたもので、宣教的意味を持つ[1]。福音書には、マルコ、マタイ、ルカ、ヨハネ福音書の四文書がある。これら四福音書のうち、マルコ、マタイ、ルカ福音書の三福音書（以降、場合によりそれぞれ Mk,Mt,Lk と略す）については、互いに密接な類縁関係があり、三つの並行するフレームからなる対観表の形にあらわすことができるため「共観福音書」と呼ばれている[2]。そして、この共観福音書を様々な共通単元のフレームで並べ換え、相互に同時比較できるようにしたものが「共観表（Synopsis）」である。これは、J.J.Griesbach が 1974 年に出版した『共観福音書対観表』[3]においてはじめて用いた言葉であり[4]、現在の新約学では、Kurt Aland が作成したギリシャ語共観表が最も信頼性のある共観表として認められている[5]。

新約聖書学においても言語の一致度の割合を計算して、共観表の切り分け問題や、想定している資料についての議論を行っている。Linnemann は、並行箇所の一致度を計算し、マタイ・ルカの共通部分から想定された Q 資料部分といわれている箇所に重点を置き、その一致度の低さから、Q 資料を否定する見解を示している[6] [7]。

2.2 フレーム問題

ここで「Q 文書」は、認知科学的にいうと、伝統共観表という「フレーム」の中で推論することを可能にする「ヒューリスティックス」であると見なしうる。すなわち、Linnemann の指摘同様、それが「錯視」ではないと言い切れる根拠はどこにも存在しない。

計算共観表は、実証的根拠から伝統共観表の不適切なフレーミング（文書の切り分け）を是正する。われわれはそこに、研究者・学習者のメンタルマップ（まさしく伝統共観表がそれに相当する）と、客観的なデータに基づく知識マップを照合し、差異を明らかにするコンセプチュアルマップ(Sadaani *et al.*)の効果[8]を認めることができる。計算共観表は、Sadaani *et al.*が提起した、「学習者（研究者）の認知構造を同定しその誤解をただす」というコンセプチュアルマップの効果をもつといえる。

2.3 TextSegmentaion

人文科学の領域においても、ヒューリスティックスにより確率統計的な方法は隅に押しやられている。単語の頻度データに、基づく統計言語学、計量文体論の応用は小規模であり十分でない。フレームの外に立ってそれを相対化するばかりでなく、フレーム自体を別の形に更新する必要がある。つまり、対象となる文献に関し、「世界の切り分け方を変える」必要がある。

どのテキスト分割 segmentation を選んでデータを取得すべきか、データフィールドの画定の問題が、一種の「フレーム問題」[9]の形で、根本に横たわっているのだ。われわれの統計言語学的な経験からすると、学習者（研究者）のヒューリスティックスは、実際にテキストを分割するフレーム（枠付け--認知的、かつ言語的な意味で）によって限界付けられている場合があるからである[10]。

このように、恣意的なテキスト分割は解釈者のもつフレーム構造と直結しており、しかもこの伝統的なフレーム構造の学習が対象の学習において大きな比重を占める。このテキスト分割(segmentation)の問題を解決しない限り、単語の頻度データをもとにした計量解析は有意義なものとなりえない。そしてそのテキスト分割(segmentation)の理論と応用は、すなわち新しい同一性の定位は、文章ベースのシソーラスにコンセプチュアルマップの利点（理解再構成を促すシャッフル）を導入することで可能になる。

3. 作成目的

われわれは、新約聖書学において、長い間討論されてきた「共観福音書問題」をめぐり、この問題を解決するために提唱された仮説について検討した[11]。

具体的には、共観福音書に出現する単語の頻度数を、異なる方法で７つの文書カテゴリーに分配する。７つの分配カテゴリーは、３福音書をベン図で表して説明すると（図１）、３書共通部分（A)、マタイ・マルコ共通部分(B)、マルコ・ルカ共通部分(C)、マタイ・ルカ共通部分（D)、と、それらの共通部分を除いたマタイ(E)、マルコ(F)、ルカ（G）部分である。

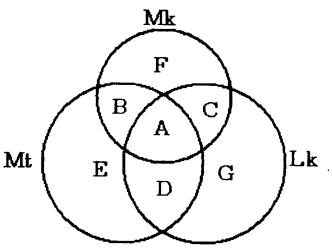

図 1：７つのカテゴリー

分析に用いた３種類のデータセットは、従来の聖書学における伝統的共観表に基づいて作成した。さらにそのデータ行列（文書カテゴリー×単語）を用いて多変量解析（因子分析）を試み、伝統的な仮説とその出力結果を比較・検討し、さらには福音書の成立の謎に迫った。

しかしながら、この分析で使用した共観表は、単語の一致という客観的な指標に対して、幾つか不適当な性質を持っている。まず、ペリコーペ（物語）単位によって切

れ分けられているため、一並行箇所における各文書の長さの相違が極端な箇所がある。そして、共観表の性質上、3 文書共通の並行箇所がその大半を占めており、マタイ・ルカ共通箇所、マルコ・ルカ共通箇所の割合が極めて少ない。さらに、文章の一致がそれほど確認されない箇所においても、内容の一致と称して恣意的に並行箇所にまとめていると考えられる箇所も見受けられる。以上のような並行箇所の切り分けの問題点が、伝統共観表に基づく分析結果に影響を与えていると考えられる。特に、その影響は、並行箇所データセットにおける、カテゴリー間の単語量の偏りによって顕著に表れている。したがって、より客観的な共観福音書の文書成立の計量的分析を行うために、単語の一致を指標にして並行箇所を切り分ける「計算共観表」を作成する必要が生じた。

4. 異本アルゴリズム（SynopticPatch）

異本解析アルゴリズムは、文章の一致度が高い 2 つのテキストから、文書の類似度を表す相関係数を参考にして、共通している部分を抽出するアルゴリズムと定義できる。計算共観表作成にあたっては、様々な異本解析アルゴリズムが考えられる。比較される文章における単語の完全な一致度を起点にするものや、文・文章の類似度をベースにするものなど、可能な選択肢は数多く存在するだろう。ここで論じる SynopticPatch は、あくまでそのうちのひとつである。SynopticPatch は、*N*-gram と Windowing と TextTiling [12]の 3 つを組み合わせた方法論である。

まず *N*-gram モデルは、確率・統計的自然言語処理の分野で使われている、強力な言語モデルである[13]。2 つの比較テキストに関して、並行対応する最長一致 *N*-gram インスタンスを見つけ、それぞれの開始位置、終了位置を記したリストのリストを用意する。そのリストにおいては、互いに対応する要素（サブリスト）、すなわち共通の（最長一致）n グラムインスタンスがもつテキスト中の位置(自然数値)を示している。一方、テキスト中の位置レベルばかりでなく単語レベルでも同様のリストを用意する。その両リストの対応関係をもとに単語情報の計算を制御する。

原理的には、並行単語列リストの要素であるサブリスト、すなわち、並行対応する最長一致 *N*-gram インスタンスを共起情報取得用ウィンドウの中心に置く。この並行対応する最長一致 *N*-gram インスタンスを「並行島」と称する。そして比較対照する双方のテキストで、同期を図りながら左右双方に対してそれぞれ共起ウィンドウを伸長してゆく。

つまり SynopticPatch は、*N*-gram インスタンスを中心に置くウィンドイング法であり、しかも 2 つのウィンドウを並行テキスト間で同時にランデブー走行させるものである。しかし、ここでは単に共起頻度を取得するばかりでなく、ウィンドウが 1 単語ずつ新たに外側の単語を拾ってゆくごとに、並行ウィンドウ内の単語頻度ベクトルの相関係数を計算する。

ここで、文書の類似度を判定する祭に、Salton のベクトル空間モデル[14]による文書の類似度に従って求めた。

ウィンドウの中心の並行島の相関係数は、1 である。しかし、ウィンドウが中心

Ngram インスタンスをいったんはみ出すと、両テキストで異なる単語を拾うので、類似度はどんどん落ちてゆく。SynopticPatch は、並行島と他の地点との間を同時並行的に埋める(継ぐ)ということから名前がつけられている。ウィンドウの伸長停止条件には様々なものがあり、どちらかの側で隣の並行島に到達したら停止するという条件も設定できる。

最大に伸長したウィンドウ内で、並行箇所の切り分けを行う際の基準に関しても様々な方法が想定でき、対象となる並行テキストの性質に最も合致したものを見つける必要がある。最も原理的な SynopticPatch 法では、Hearst の TextTiling に習い、相関係数の落ち込みを「深度」として、並行島の海面下で並行島がそこから屹立する地点の深度をもって、並行箇所を自動決定する。

このようにして、類似した2つの文書から共通部分をとりだす。図2は一般的な意味での SynopticPatch を図解したものだが、共観福音書の並行箇所処理は、この手法をさらに文書の実情に合うようにカスタマイズして適用した。これについては次節で詳述する。

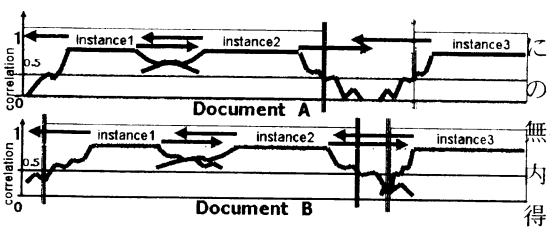

図 2：Synoptic Patch

5. 作成方法

異本アルゴリズム (SynopticPatch) を適用した、計算共観表の作成方法について述べる。今回の計算共観表の作成における第一目標は、伝統型共観表の並行箇所の偏りを減少することである。計算共観表は、2つの伝統共観表 (Q:Mt-Lk, Mt-Lk-Mk) における並行箇所の切り分け方に関する、切り分けの精度、細かさの差（意味上の密度、濃密さの差）を調整することによって得られる。Q は、人々の関心が集中しているので、存在を立証しようとする努力もあって、切り分け方が細かい。一方三つ揃い Mt-Lk-Mk は概括的なものであって、切り分け方が粗い。

したがって、伝統型共観表を教師並行箇所としながら、計算の対象とする文書範囲を定め、各教師並行箇所における文書の比較・切り分けを行った。

また、SynopticPatch は類似した 2 文書間の共通部分の切り分ける方法論である。これを 3 文書の共観福音書に応用するために、マタイ・マルコ、マタイ・ルカ、マルコ・ルカといった 2 文書間から求められる各文書における並行島の位置で、一致する部分を 3 文書共通部分の並行箇所とした。

ここで、並行島とする Ngram の大きさについては、最小値を3と定め、それ以下の共通箇所、すなわち bi-gram については無視した。一方、並行島を構成する単語列内の最大単語数であるが、設定並行箇所で得られた最大共通数を最大値とした。一般的に共観表とは、そこに単に意味や文脈の深い類似性ばかりでなく、作者間の明確な「引用関係」の可能性が認められて、初めてその「単位」としての並行箇所が規定で

きるはずのものである。そうなると、2 語の単語からなる bi-gram のインスタンスは、マイナーアグリーメントの場合を除き、引用に基づく並行島として弁別しにくい。また bi-gram のインスタンスは多数存在し、その間で並行対応関係をつけるのは困難である。また、設定した並行箇所が広範囲の時、あるいは Nの値が小さい場合には、一並行箇所内で複数の同単語列が出現し、比較する文書間で得られる並行島が一意に定まらないことがある。この場合は、両文書の並行島の数が一致するまで、N の最小値を大きくしていき、並行島の補正を行った。

さらに、並行島を中心としたウィンドウィングの幅については、設定並行箇所の両端まで伸張させていき、その範囲で文書間の相関を計算した。この方法は、並行島の出現順序が文書間において異なる場合を想定している。そして、並行箇所の内部で複数の並行島がある場合、その並行島ごとに、並行箇所全体を覆う同じ長さのウィンドウが（並行島の数だけ）「重なって」できることになり、それらの系列をここで「地質断面」と称する。

ウィンドウの中心においてはむろん相関係数（コサイン類似度）は 1 である。しかし、ウィンドウが中心 n グラムインスタンスをいったんはみ出すと、両テキストで異なる単語を拾うので、類似度はどんどん落ちてゆく。単語を拾うごとにコサイン類似度を計算しベクトル型のリストに記録する。SynopticPatch は、並行島と他の地点との間を同時並行的に埋める（継ぐ）ということから名前がつけられている。最終的に複数の「地質断面」の相関係数を平均した値を、文書の相関係数とし、経験的に閾値を 0.5 と定め、節区切りで文書の切り分けを行った。むろん、この切り分け基準も様々な方法（たとえば、平均を取らず、最左、最右の地質断面のみの推移を見るものなど）が考えられるうちのひとつである。

このようにして、SynopticPatch のカスタマイズを行い、並行箇所を計算して求めた。具体的には以下の方法によった。

共観福音書の中から、具体的には伝統型共観表より、総計 143 個の設定並行箇所に対し、個別に SynopticPatch の計算を行った。表 1 に使用した並行箇所のカテゴリー別の個数を示す。

A	B	C	D	合計
101	6	3	33	143

表 1: 計算に使用した並行箇所数

それぞれの並行箇所において、まず先述の原則に則って、3 文書間の一致部分を抽出し、Mk-Lk, Lk-Mt, Mk-Mt の２つ組にそれぞれ振り分けてから、並行ウィンドウ間の相関を取った。

6. 計算共観表の特徴

伝統共観表、計算共観表それぞれにおけるカテゴリー別の出現単語数を表 2 に示す。

また、伝統共観表、計算共観表それぞれにおけるカテゴリー別の節数の割合を表 3, 4 に示す。

	伝統	計算
A	4877	3781
B	606	1021
C	244	478
D	1915	1623
E	1061	1838
F	75	747
G	2015	2762

表 2：カテゴリー別出現単語数

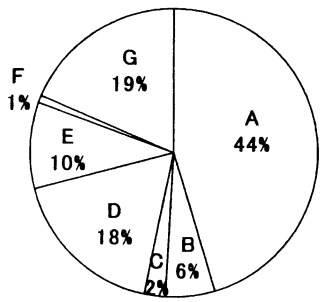

表 3：従来共観表における節数の割合

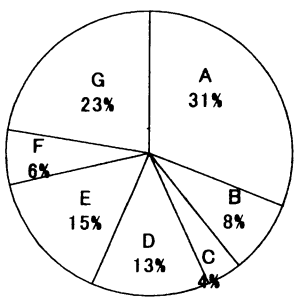

表 4：計算共観表における節数の割合

表 3 で、3 文書共通カテゴリーA が 42 パーセントを占めていたのに対して、表 4 では、その約半分の割合に減少している。これは、一定度の単語の一致が見られない節が削除され、独自カテゴリー（E,F,G）に移動したことが考えられる。また、カテゴリーB、C の増加から、カテゴリーA の箇所であった一部が、他の共通カテゴリー（B,C）へ移行したと考えられる。

さらに、両者の共通カテゴリー（A+B+C+D）の割合を比較すると、表 5 に示す結果となり、

	伝統	計算
割合	60％	42％

表 5：共通カテゴリーの割合

この結果から、計算共観表は各文書の独自カテゴリーの増加が著しいことが分かる。この理由としては、SynopticPatch の最初の手順である、3 単語以上の共通一致単語の並行島を求める作業において、幾つかの伝統型共観表においては並行箇所であった部分がなくなり、各文書の独自部分へと移行したことが考えられる。一例としては、並行箇所番号 6, 7, 8, 11, 25 が挙げられる。

7. まとめと今後の課題

SynopticPatch を共観福音書に適用する場合、以下のような問題点が生ずる。
並行比較されるふたつの福音書問で

- 隣接する並行島インスタンスの位置が前後入れ替わり、ウィンドイングが交絡を作ることがある
- 同一の並行島インスタンスが他方で重複し、ウィンドウが分岐してしまう場合がある。

交絡と分岐の問題があるため、SynopticPatch をそのままの形で適用する

ことができなかった。また今回はQ文書を中心に考えてマイナーアグリーメントの部分が抽出できなかったが、それは、並行島の条件から trigram 以上に絞り、bi-gram を省いたことによる。Q の場合とちがい、多くの研究においてマイナーアグリーメントの根拠付けには、いくつかの bi-gram のインスタンスが現に使用されている。しかし、bi-gram の出現確率の高さを考えると、bi-gram のインスタンスをすべて無条件に使用するのは、交絡・分岐の処理の複雑さからコスト的にも、さらには希少性に基づく存在価値の点で理論的にも、不適切といわざるをえないであろう。bi-gram インスタンスの取捨選択に関する条件設定が今後の課題として残る。

8. 謝辞

本研究は、21 世紀 COE プログラム(研究拠点形成補助金)「大規模知識資源の体系化と活用基盤構築」の言語・文献、知識資源分野に関する研究の一環として行われたものである。また、Tele-COEX の開発にあたって、Mathematica のアドバイスをいただいた東工大学術国際情報センターの松田裕幸先生に感謝します。

【参考文献】

[1]. Conzelmann, H. & Lindemann, A., Interpreting The New Testament, trans. by Siegfried S. Schatzmann, Hendrickson Publishes,45-53, 1988.

[2]. Theissen, G., *Das Neue Testament*, Beck, Mchn, 2002.

[3]. Synopis Evangeliorum Matthaei, Marci et Lucae, Greisbach, Helle, 1776

[4]. Kloppenborg, John S., et al. *Q Thomas Reader* , Polebridge Press, 1990

[5]. Nestle-Aland, *Novum Testmentum Graece 26th edition*, German Bible Society Stuttgart

[6]. E. Linnemann, Is There a Synoptic Problem?, Grand Rapids:Baker, 1992.

[7]. E. Linnemann, "The Lost Gospel Of Q-Fact Or Fantasy",Trinity Journal. 17:1, pp.3-18, 1996

[8]. Lalthoum Saadani & Suzanne Bertrand-Gastaldy, "Conceptual Maps and Thesauri: A Comparison of Two Models of Representation from Different Disciplinary Traditions", *CAIS 2000: Dimensions of a Global Information Science*, 2000

[9]. Minsky, M.L., A Framework for representing knowledge, The Psychology of computer vision, pp.211-277, 1975

[10]. Matsubara, "Frame-mondai no Giji-kaiketsu no tame no Heuristics toshite no Ingaritsu", *Ninti-kagaku no Hatten*, vol.6, 1990

[11]. 三宅真紀, 赤間啓之, 佐藤研, 中川正宣, 使用単語の因子得点に基づく福音書ジャンルの特徴考察, *文理シナジー学会平成 16 年度大会発表要旨集*, p.18, 2004

[12]. Hearst, Marti A., TextTiling: "Segmenting text into multi-paragraph subtopic passages", *Computational Linguistics* 23, pp.33-64, 1997

[13]. 北研二、確率的言語モデル、東京大学出版会、1999

[14]. — E Salton, G., "A theory of Indexing", Society for Industrial and applied mathematics, 1975