E-5

多言語特許検索システムの開発と評価

牧田 光晴 †, 樋口 重人 †, 藤井 敦 ††,††† 石川 徹也 ††

†(株) パトリス

†† 図書館情報大学

††† 科学技術振興事業団 CREST

1 はじめに

近年，特許検索の領域で海外特許を検索することが重要になってきており，筆者らは多言語特許検索システム **PRIME**（Patent Retrieval In Multi-lingual Environment）を開発した [2, 7]．PRIME は検索キーワードを検索対象の言語に翻訳して外国語特許を検索し，さらに検索された特許をユーザの母国語に翻訳することで多言語検索を実現する．また，検索結果を分類（クラスタリング）することで，閲覧効率を向上させることができる．

翻訳や検索については既に評価実験を行い，概ね良好な結果を得ている．本稿では，主に，クラスタリングの性能に関する評価を行う．

2 PRIMEのシステム構成

PRIME のシステム構成を図 1 に示す．実線はオンライン処理，破線はオフライン処理をそれぞれ表す．現在，日本語と英語による検索が可能である．検索対象データは日本特許抄録 1995-1999 年の 5 年分（約 175 万件）[1]，英語抄録 PAJ（Patent Abstracts of Japan）[2] の同期間，同件数含むパラレルコーパスである．

ユーザがキーワードやフレーズなどで検索質問を入力すると，翻訳部において目的言語に翻訳される．ここでは，対訳辞書を引き，単語や複合語の単位での翻訳を実現する．しかし，一つの語に対し複数の訳が存在し，訳語候補の全てを採用するとノイズが増大するため，対訳辞書から抽出した翻訳モデルと検索対象の特許 DB から抽出した言語モデル（単語バイグラム）を使用して統計的手法で翻訳の曖昧性を解消する．我々の予備実験 [7] では，翻訳の精度は，異表記や省略等による表記の違いも許容すれば，日英で 94.2 %，英日で 92.8 %であり，実用レベルに達している．

検索質問と特許の類似度を確率的なスコアで計算し，スコアに基づいて検索された特許をソートしてユーザに提示する．この段階で，検索結果をクラスタリングして，クラスタ（グループ）の単位でユーザに特許を提示することが可能である．その結果，個々の特許ではなく，クラスタ単位で，必要な情報を選択することができる．さらに，この処理を対話的に繰り返しながら，ユーザを適切な情報に効率良く誘導することができる．

特許文書を検索，分類する際には，国際特許分類（IPC）などが使われる．しかし，分類体系を把握せずにこれらを駆使することは困難である．そこで，特許の内容に基づいてクラスタリングを行う．この方式であれば，IPC などに関する知識を必要とせずに分類が可能になる．

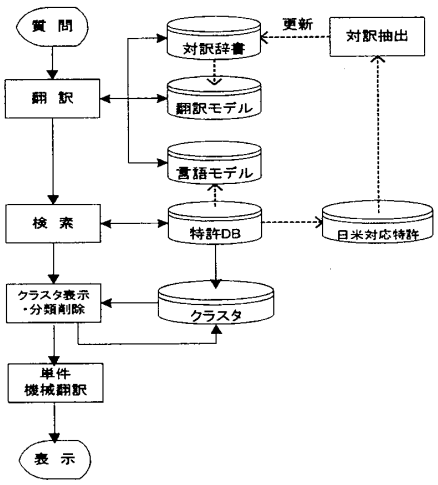

図 1: PRIME のシステム構成

最後に，ユーザがクラスタから特許を単件指定すると，その内容がユーザ言語に機械翻訳されて表示される．

また，オフライン処理ではパラレルコーパスから日英対訳を自動抽出し，対訳辞書を随時更新する [8]．

3 PRIMEの評価

ここでは，筆者らの先行研究で評価の対象となっていなかったクラスタリングの性能について評価を行う．

クラスタリングの評価には，いくつかの方法がある．Hearst ら [3] は，まず検索結果をクラスタリングして作成された複数のクラスタから，唯一の「最適な」クラスタを決定し，適合文書（正解）が最適クラスタに含まれる確率によって，クラスタリングの精度を評価した．

Fujii ら [1] は，クラスタリングの評価にエントロピーの概念を導入した．すなわち，特定のクラスタに適合文書が集中し，適合文書の分布に偏りが生じるほど，良いクラスタリング手法と見なす．逆にどのクラスタにも適合文書が一様に分布する場合には，ユーザが適合文書を選択するための情報がほとんどないので，クラスタリングの精度は低いと見なす．

[1] (株) パトリスより販売
[2] 特許庁より販売

表 1: クラスタリングの比較評価結果

アルゴリズム	文書数	JE	JJ
クラスタリング無し	100	0.895	1.47
	500	0.538	0.832
	1000	0.408	0.624
HBC	100	0.752	1.31
	500	0.506	0.841
	1000	0.422	0.632
SLINK	100	0.753	1.15
	500	0.360	0.715
	1000	0.168	0.409

Fujii らの手法に基づき，式（1）で示されるエントロピーを用いて，クラスタリングをすることによりどれだけ適合文書を発見することが容易になったかを定量的に求めるための評価実験を行った．

$$H(X) = -\sum_{c} P(c) \log P(c) \qquad (1)$$

ここで，$P(c)$ は各クラスタ中に適合文書が含まれる確率を表し，実際には，式（2）を用いた．エントロピーが小さいほど，クラスタリングによってユーザが適合文書を発見するのが容易になったと判断する．

$$P(c) = \frac{\text{クラスタ中の適合文書数}}{\text{クラスタ中の総文書数}} \qquad (2)$$

本研究の評価には，NTCIR-3 の特許タスク [6] で配布されたパトリス検索課題 34 件を使用した．また，クラスタリング手法には，一般的な SLINK （Single Linkage Method：単一リンク法）[4] と，確率的階層クラスタリングである HBC（Hieralchical Bayesian Clustering）[5] の 2 つを用いた．HBC では，ほぼ均等なサイズのクラスタが作成されるのに対して，SLINK では，作成されるクラスタのサイズは不均一である．さらに，SLINK では，検索結果から完全に異種なものを切り離してゆく「慎重な」クラスタリングが行われるため，適合文書が多く含まれるクラスタが他のクラスタよりも大きくなりやすい．

また，検索件数の最大値を 100 件，500 件，1000 件の 3 通りに変えて行った．表 2 に 34 の検索質問について，アルゴリズム，検索方法別に行った実験結果（エントロピー）の平均値を示す．表 2 において，JE は日英検索，JJ は日本語どうしの単言語検索を表す．JE では日本語の検索質問に翻訳を施した後に PAJ を検索し英語データの検索結果にクラスタリングを施したものを日本語の正解集合と照合した．この場合，日本語抄録と PAJ の文献番号は同一であるため，テストコレクションをそのまま用いることができる．一方，JJ では日本語の検索質問で日本語データを検索・クラスタリングした結果をテストコレクションと照合した場合をそれぞれ表す．また，クラスタリング無しの場合は，検索結果をクラスタリングの場合と同様の文書数でカットオフし，それをランク順に 5 個の群に分けたものをクラスタと見立て，同様な手法でエントロピーを算出した．

JE では，クラスタリング無しの場合と比較して，HBC，SLINK いずれのアルゴリズムにおいても，クラスタリングによるエントロピーの減少が見られた．例えば，文書数 100 件の場合，クラスタリング無しとくらべて SLINK ではエントロピーが 0.143 低下しており，同様に文書数 1000 件の場合ではエントロピーが 0.240 低下がっている．

また JJ では，SLINK の場合はいずれもエントロピーは軽減し，HBC でも検索結果集合が比較的小さい場合にはこれが減少した．これらのことは，クラスタリングによりユーザの負荷が減少することを示唆するものと考えられる．また，すべての場合において JE と JJ を比較すると，全体的に JE のエントロピーの値の方が小さいことから，日本語と英語では後者の方がうまくクラスタリングされていると考えられる．

本研究では，いずれの評価実験も検索結果を一度クラスタリングしたものに対してしか行われていない．本来のクラスタリングでは，反復的にクラスタリングを行いながら文書群をしぼるため，ここで行った評価に加え複数回のクラスタリングの評価も求められるだろう．

アルゴリズムの比較では，SLINK は常に HBC を上回る結果が出た．しかし，一回のクラスタリングの評価であるため，また，先に触れたように，HBC と SLINK とでは作成されるクラスタサイズに大きな違いがあるため，反復クラスタリングの評価を行えばこれらの結果が変わる可能性も残されている．

4 おわりに

PRIME の機能として，母国語による質問入力，質問翻訳，他国語データの検索，一覧表示，検索結果群のクラスタリングによる分類，検索結果単件の機械翻訳を実現した．また，評価実験によってクラスタリングの有効性も確認できた．今後は，反復的クラスタリングの評価や，韓国語をはじめ，対応言語の拡張を行う予定である．

謝辞 本研究では（株）ノヴァの許諾を得て，専門用語辞書および特許機械翻訳システム PatTranser を使用させて頂きました．この場を借りて深謝致します．

参考文献

[1] Atsushi Fujii, Tetsuya Ishikawa. Evaluating multi-lingual information retrieval and clustering at ULIS. *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 2001.

[2] Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. PRIME: A System for Multi-lingual Patent Retrieval. *Proceedings of MT Summit VIII*, pp.163–167, 2001.

[3] Marti A. Hearst, Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. *Proceedings of ACM SIGIR*. pp.76-84, 1996.

[4] Peter Willet. Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24(**5**), pp.577–597, 1996.

[5] 岩山真，徳永健伸. 確率的クラスタリングを用いた文書連想検索. 自然言語処理，**5**(1). pp.101–117, 1998.

[6] 岩山真，藤井敦，高野明彦，神門典子. 特許コーパスを用いた検索タスクの提案. 情報処理学会研究報告 2001-FI-63, pp.49–56, 2001.

[7] 樋口重人，牧田光晴，藤井敦，石川徹也. 多言語特許検索システム PRIME. 言語処理学会第 8 回年次大会発表論文集, pp.196–199, 2002.a

[8] 福井雅敏，樋口重人，藤井敦，石川徹也. 日米対応特許コーパスを用いた対訳抽出手法. 情報処理学会研究報告 2001-NL-145, pp.23–28, 2001.