2G2-OS9-1

最小ユーザフィードバックによるインタラクティブ情報収集・データマイニングの枠組み

Interactive Information Gathering and Data Mining with Minimal User Feedback

山田 誠二[*1] 小野田 崇[*2] 高間 康史[*3] 岡部 正幸[*4]
Seiji Yamada Takashi Onoda Yasufumi Takama Masayuki Okabe

[*1]国立情報学研究所／総合研究大学院大学／東京工業大学 [*2]電力中央研究所
National Institute of Informatics/SOKENDAI/Tokyo Institute of Technology CRIEPI

[*3]首都大学東京 [*4]豊橋技術科学大学
Tokyo Metropolitan University Toyohashi University of Technology

In this paper, we propose a framework for interactive information gathering and data mining in which a user is able to gather information with minimal user feedback. In order to such an intelligent interactive system, we are developing constrained clustering with minimal user's constrains, clustering algorithms with Independent component analysis-based seeding and GUI for relevance judgment with less cognitive load. By combining these fundamental technologies, information gathering and data mining with minimal user feedback will be achieved.

1. はじめに

現在，一般ユーザがインターネットを介して自由に利用できる Web，ブログなどのテキストデータは，かつて人類が経験したことのない程の大規模な情報源になっている．このような情報を有効に利用する方法論の確立は，情報工学，情報学に課せられた急務であることは誰しも認識している．しかし，残念ながら情報検索の専門知識を有しない一般ユーザが満足のいく程度に効率よく欲しい情報を収集する情報収集・データマイニングシステムの構築は，未だ困難な課題である．

本稿では，このような問題に対処するため，ユーザの負担を最小限に抑えたユーザフィードバックにより最大限の効果を引き出す，最小ユーザフィードバックによるインタラクティブ情報収集・データマイニングについて，その枠組み，必要となる要素技術を検討，考察する．

2. インタラクティブ情報収集・データマイニング

一般ユーザが満足のいく，大規模データからの情報収集・データマイニングシステムを実現するためには，システム単独の能力向上では限界があるため，ユーザに簡単で有効な支援をしてもらい，人間とシステムが協調作業を行うことが重要である．このような背景から，我々が最も有望と考える情報収集の枠組みが，インタラクティブ情報収集・データマイニング（図 1）である．インタラクティブ情報収集・データマイニングとは，ユーザがシステムを助けながら，協調して情報収集・データマイニングを行うメカニズムである [Onoda 07]．具体的には，初期制約としてクラスタリングの制約や検索クエリをシステムに入力し，システムの返してきた結果であるヒットリストやクラスターをユーザが評価することで，新たな制約をユーザフィードバックとしてシステムに与える．その制約に基づき，システムが再検索，再クラスタリングをを行い，よりよい結果を再度ユーザに提示する．このループの繰り返しにより，ユーザ所望の結果が得られる．

連絡先: 山田誠二，国立情報学研究所／総合研究大学院大学／東京工業大学，東京都千代田区一ツ橋 2-1-2，seiji@nii.ac.jp

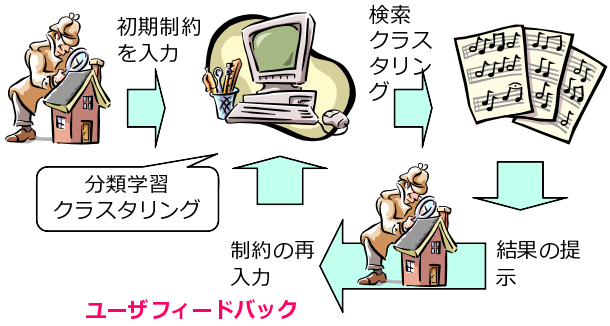

図 1　インタラクティブ情報収集・データマイニング

しかし，この枠組みには克服すべき問題がある．まず，ユーザフィードバックと呼ばれる，ユーザによる評価自体が，ユーザにとって大きな負担となることである．例えば，通常の文書検索では，ヒットリスト上位の 20〜40 程度の文書の判定が必要となるが，その判定のためには，それらの文書を精読する必要があり，このユーザフィードバックがユーザにとって多大な認知的負担となる．よって，システム全体のパフォーマンスを落とさずに，このユーザフィードバックをいかに小さく抑えるかが，この枠組みの成功にとって必須条件となる．

以上の背景から我々は，従来のシステムのパフォーマンスを維持しつつ，ユーザフィードバックを最小にする最小ユーザフィードバックによるインタラクティブ情報収集・データマイニングの枠組みを提案し，その実現に必要な要素技術を開発する．

3. 最小ユーザフィードバック

本研究の目的は，従来の情報収集・データマイニングシステムと同等かそれ以上のパフォーマンスをもち，かつユーザフィードバックを最小にする最小ユーザフィードバックによるインタラクティブ情報収集・データマイニングの実現である．すなわち，図 2 のように，システム全体のパフォーマンスを維持しつつ（図 2 の $P \leq P'$），ユーザフィードバックを最小に

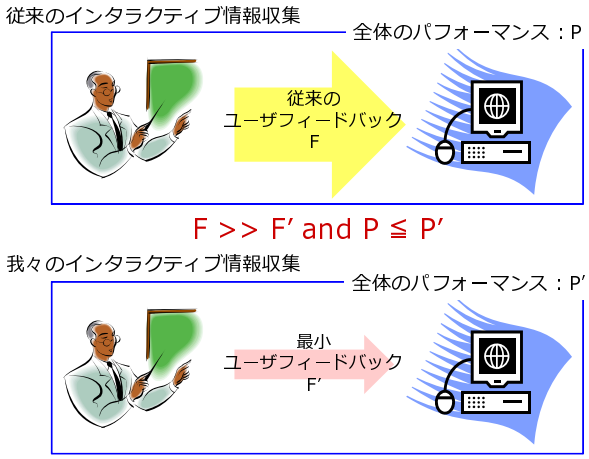

図 2　最小ユーザフィードバックのコンセプト

する（図 2 の $F >> F'$）ことを目指す．

まず，ユーザフィードバックを計算論的，認知的の両側面から捉える．計算論的最小化は，少数の制約でも十分に精度の高い制約クラスタリングアルゴリズムの開発により実現する．この技術により，制約の数をできるだけ少なく抑えることができる．しかし，個々の制約を人間に与えてもらう認知的負荷が高いと意味がない．よって，できるだけ認知的負荷を少なく，人間のユーザが個々の制約を与えることのできる GUI を併せて開発する．これらの技術により，制約数を最小限に抑え，個々の制約獲得の認知的付加を最小限に抑えることができ，最小ユーザフィードバックが実現される．

このような考えによる最小ユーザフィードバックによるインタラクティブ情報収集・データマイニングは，これまでに例を見ない独創的なものと考えられる．制約クラスタリング [Basu 08]，少数制約とクラスタリングの関係を調べた研究 [Iwayama 00] などの研究が関連するが，ユーザフィードバックを最小限に抑える具体的な方法論を開発していない．ユーザフィードバックを極力抑える視点もないし，最小ユーザフィードバックの定義もない．さらに，情報検索において認知的負荷を考慮した枠組みも前例をみない．

4.　開発中の要素技術

最小ユーザフィードバックを実現するため，現在開発中の要素技術を以下に示す．ここでは，最小ユーザフィードバックを適用する情報収集・データマイニングとして，広く使われているデータマイニング手法である制約クラスタリング [Basu 08] を採用する．制約クラスタリングとは，（ユーザによって）与えられた制約をできるだけ満たすようにするクラスタリングである．典型的な制約は，同じクラスタに属するべきデータペアである Must-link と別のクラスタに属するべきデータペアの Cannot-link があり，本研究でもこれらの制約を扱う．

4.1　少数制約下の制約クラスタリング

少数制約でできるだけ精度のよい制約クラスタリングアルゴリズムの開発を行っている [Okabe 10]．与えられた制約を近傍のデータにも伝播することで，少数制約を効率的に利用する拡張も行っている [Okabe 10]．基本的には，グラフカットベースの距離カーネル学習アルゴリズムを用い，さらに能動学習を組み込む予定である．

4.2　独立成分分析による制約クラスタリング

独立成分分析により得られた独立成分のベクトルを従来の非階層型クラスタリング手法である k-means 法のシードに利用する．複数のシードにより何度もクラスタリングを繰り返す必要なく，一度の試行で高精度のクラスタリングを実現する．

独立成分分析により得られる複数のソースベクトルがほぼ同じ方向を持つ場合は，そのときは K-means++[Arthur 07] と類似の方法でシードとなるデータを決定する．

4.3　認知的付加の低い類似文書判定の GUI

制約クラスタリングの制約は，基本的に 2 つのデータの類似度に基づき判定される．そのため，2 つの文書データの類似度を判定しやすい，つまり認知的付加のかからない文書の類似度判定用に特化した GUI の開発を行う．

具体的には，テキストそのまま，スニペット，特徴的なタームのリストなどを提示するインタフェースによる判定効率の違いを調べ，認知的付加の少ない類似文書判定の GUI を設計する．また，グラフベースの GUI[Takama 07] などの新しい提示方法を開発する．

5.　まとめ

できるだけユーザの負担にならないユーザフィードバックにより最大限の効果を引き出す，最小ユーザフィードバックによるインタラクティブ情報収集・データマイニングについて，その枠組み，必要となる要素技術を検討，考察した．今後，要素技術の開発と実験的評価と平行に，システム全体のまとめを行っていく予定である．

参考文献

[Arthur 07] Arthur, D. and Vassilvitskii, S.: k-means++: the advantages of careful seeding, in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035 (2007)

[Basu 08] Basu, S., Davidson, I., and Wagstaff, K. eds.: *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Chapman & Hall (2008)

[Iwayama 00] Iwayama, M.: Relevance feedback with a small number of relevance judgements: incremental relevance feedback vs. document clustering, in *In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 10–16 (2000)

[Okabe 10] Okabe, M. and Yamada, S.: Learning Similarity Matrix from Constraints of Relational Neighbors, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 14, No. 4, p. to be published (2010)

[Onoda 07] Onoda, T., Murata, H., and Yamada, S.: SVM-based Interactive document Retrieval with Active Learning, *New Generation Computing*, Vol. 26, No. 1, pp. 49–61 (2007)

[Takama 07] Takama, Y., Matsumura, A., and Kajinami, T.: Interactive Visualization of News Distribution in Blog Space, *New Generation Computing*, Vol. 26, No. 1, pp. 23–38 (2007)