D-40　活動データのマイニングによるデータ優先度の決定

Computation of Data Priority by Mining of Activity Data

成凱†　平野真太郎‡　李龍†　東郁雄⊥　上林弥彦†
Kai Cheng　Shintaro Hirano　Ryong Lee　Ikuo Azuma　Yahiko Kambayashi

1. 概要

データを広域に分散して保持している会社では各地域ごとにデータサーバーを設置してデータベース作成管理者が割り当てられている。この形では不要なデータ除去や優先順位決定は各データ管理者が決めている。しかしながら社内システムを統合化しデータを有効利用するためには、これらのデータを集中管理する必要がある。統合データに対しては、どのようなデータが優先順位が高いかを決定することが非常に困難であるが重要であると言える。従来のデータの優先順位を決める最も簡単な方法は LRU であり、データの容量に上限があり新しく追加されるデータがデータベースに追加される場合、古くから存在していて使われていないデータが追い出されて新しいデータが加えられると言う形であった。しかし半年おきに使われるデータなど長期的に見たときに利用されるデータが三次記憶に出されているとデータ利用の効率が低下することが考えられる。

このため統合されたデータに対して優先順位を決める方法を決定することが非常に重要となる。ここでは３つの方法を併用して優先順位を決める方法を提案する。

- 各データに対するアクセスの頻度や最終的に利用されてからの時間だけでなくより知的な方法により優先順位を決める。具体的には必要なデータに含まれるキーワードなどを用いた優先順位決定を行う。
- 上記のキーワードの決定に対しては、データベースの利用だけではなく社内の運用データを解析しデータマイニングにより重要なキーワードおよびそれに深く関連するキーワードを抽出する。
- データの性質としてどの役職の誰が書いた物か、決裁はどの役職まで行われたかに注目することで優先順位を決める重みとすることが出来る。これはデータの優先順位を決める際にユーザーが平等であるという観点で行うウェブキャッシュとは違いユーザーを役職によって分けるという、不平等性を利用することになる。

この方法によって従来のキャッシュや WEB キャッシュよりもより目的にそくした優先順位を決めることができる。

2. 基本的考察

ここで扱う問題はコンピュータシステムやウェブで扱われていたキャッシュと類似点も多い。まずこれらのキャッシュシステムについて整理する。コンピュータの CPU のキャッシュの場合はアルゴリズムが複雑であっては困るの LRU が採用されている。ウェブキャッシュではより複雑な計算をすることが可能となるため、次のような条件が考えられる。（１）最新利用性 LRU の利用。（２）長期間における利用頻度、例えばある期間は利用されて無いが特定の期間に非常に利用されているので平均すると利用頻度が高いもの。会社の決算期が例となる。（３）データの大きさ。より大きいデータが入ると小さいデータをいくつか消してしまうことになるので大きいデータに対してはそれなりに優先順位を下げる必要がでてくる。

従来のキャッシュ法の欠点は新しいデータが追加された時に初めて最も優先順位の低いデータが計算検出されそれが消去される。すなわちデータは最も優先順位が低いことが分かってから消されることになるのでこのような方法がかならずしも我々の目的に必ずしも適合しているわけではない。あるデータが重要である時それと関連するデータがあまり利用されていなくても付随して重要になるといった場合である。このため我々は重要度の高いデータに対してキーワード検出を行いそのキーワードに重みをつけそのキーワードの重みによってさらにその他のデータの重要度を推定する方法を採用する。

［仮定するデータの形式］

議論を簡単にするために、会社のデータベースの実体集合として以下のデータベースモデルを考える。

- **社員(id,名前、役職)**

これは社員のデータである。役職は会長・社長、取締役、部長、課長、それ以下の５段階で区別する。

- **文書(d-id、タイトル、作者、製作日時、サイズ、種類、レベル)**

文書の性質を現すものとして ID、タイトル、作者、製作日時、サイズ、種類、レベルをメンバに持つ。サイズは KB 単位で表示される。種類には企画書、報告書、メール、会議書類、マル秘書類などがありレベルは決裁がどの役職によって行われたかを示し、社員の役職と同様に５つ段階の区別がある。さらに上記の二つの実体集合の関連集合として、誰がどの文書を参照、利用したかを示す

- **アクセス履歴(id, 利用日時 ,d-id)**

を準備する。文書としてデータベースに保存されるデータは完成したデータであり編集中のデータは扱わない。

3. キーワードによる優先順位の決定

特定の組織の中では特定の目的をもったデータがよく利用される。それは書類自身の性格やその書類の持つキーワードが関連している。例えば、あるキーワードを持つ書類が非常によく使われているとすればあまり利用されていなくても同じようなキーワードを持つ書類はある場合には利用される可能性がある。したがってキーワードの重み付けを行いそのキーワードによって各書類の優先順位が決まるといったモデルを提案する。

［キーワードの優先順位の決定］

一定期間データベースの中に滞在したデータに付いてその滞在時間、利用頻度によって優先順位を決める。それぞ

†京都大学大学院情報学研究科
‡京都大学工学部情報学科
⊥関西電力総合技術研究所

れの書類に含まれるキーワードについて優先順位の高い書類に含まれるキーワードは高い優先順位を扶養する。これによってデータの滞在時間からキーワードの優先順位が導かれる。

［書類の優先順位の決定］

キーワードの優先順位が決まるとそのキーワードをどの程度含むかによって書類の優先順位を決めることができる。これは書類の優先順位が決まる事によりキーワードの優先順位、書類の優先順位がサイクル的に決まる。利用状況において動的にキーワードの優先順位が変わることになり、これは利用状況が変化すると優先順位が動的に変化できることを示している。

これを実現するアルゴリズムとしてはデータベースの操作よりアクセス履歴から文書の頻度 Fi がわかり、これと作成日時から現在までの時間 Ri と文書のサイズ Si より文書の優先度を求める式が得られる。

文書 i の優先度 = Fi / Si * Ri

Si は単位を KB とし 10GB を超えるデータは 10GB として扱う。Ri は単位を日とし 1 年を超えるものは 1 年として扱う。それぞれの変数の範囲は兼ね合いを見つける必要がある。キーワード抽出を ChaSen を用いて行い、それを基に Namazu のように TF/IDF 法を用いて文書とキーワードの関連度を計算する。これにより文書の優先順位を重みとして加えることでキーワードの優先順位が求められ、次にこのキーワードの優先順位を基に文書の優先順位を TF/IDF 法を用いて求めることでキーワードと文書の動的な優先順位の計算が可能となる。

4. 運用データによる重要度の決定

前節では書類の重要度を実際に蓄えられている滞在時間、利用頻度から決定した。ただし会社などではそのときに非常に重要なデータと言うのは蓄えられているデータよりも実際に社内で流れている情報やメール、会議のテーマによって決められる。したがって会議や議論の内容などを分析することによって重要度のあるキーワード、あるものをさらに重要度の高いものを抽出してそれらのキーワードの重み、優先度を高める事が考えられる。これらに関しては文書の種類 Ti、レベル Li、作者 Wi を先の優先度導出計算式に加えることで考慮ができる。

さらにデータマイニングの手法を用いると、例えばキーワード A が現れるとキーワード B も現れる確率が高いといったことがわかる。A だけでなく B を含む書類についても優先度を高めるといった作業が必要となる場合がる。

［データマイニングのための **Apriori Algorithm**］

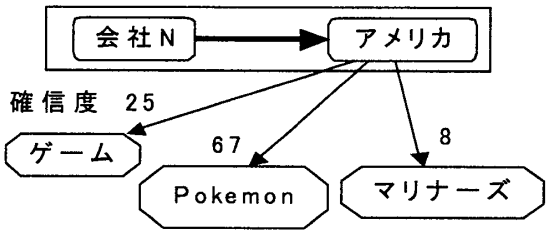
相関ルールとは，アイテムセット X(⊆I)と Y(⊆I)により，X→Y の形で記述される関係である．ここで X と Y は X∩Y=φ である． X→Y の関係は``トランザクションがアイテムセット X を含むならば，アイテムセット Y も含む"ということを表す。あるアイテムセット X について，D の内の s%のトランザクションが X を含むとき，アイテムセット X は s のサポート値(support)を持つという。また，相関ルール X→Y については， アイテムセット X∩Y のサポート値を相関ルール X→Y のサポート値と定義する．また相関ルール X→Y について，X を含むトランザクションの内の c%のトランザクションが Y も含むとき，相関ルール X→Y は c の確信度（confidence）を持つという。

相関ルールのサポート値・確信度を求めることは，アイテムセットのサポート値を求めることに集約される。 アイテムセットのサポート値を算出するとき，考えられる全てのアイテムセット数はアイテム数 N に対して 2^N 個という莫大な数になる。これら 2^N 個の全てのアイテムセットに対してそれぞれサポート値を求めるのは非常に計算コストがかかる。Apriori Algorithm は，ユーザーに指定された最小サポート値を満たすアイテムセット（頻出アイテムセット）を全て抽出する効率の良い手法で、他の多くの頻出アイテムセット抽出手法の基となっている手法である。アプリオリアルゴリズムは，あるアイテムセット X について，Y⊆X なる Y が全て頻出アイテムセットでなければ X は頻出アイテムセットではない，という原理に基づき，サポート値を数えるアイテムセットを限定する。

この場合にサポート値と確信度によってどのように優先順位を決めるかと言うアルゴリズムを以下に示す。

サポート値が高いほどデータ全体におけるその相関ルールの出現度が高いことを占めす一方、確信度が高いと言うことは相関ルールの結びつきの強さを示すものである。ゆえに今回の利用法はキーワードによって優先順位が既に決まっているものに対しての補強が目的であるので、サポート値を低くし確信度を高くすることでキーワードによる優先度順位では順位が低くても重要であるデータが発見できる。サポート値と確信度を適当な値をとりそれによって得られた注目すべきデータの優先度を確信度の大きさでもって重みとする。Figure1 では「会社 N」と「アメリカ」という優先度の高いキーワードで Apriori Algorithm を用いて得られたキーワード(ゲーム、Pokemon、マリナーズ)を現すイメージ例である。

Figure1.相関ルールから得た新しいキーワードの発見例



この新しく得られたキーワードは「会社 N、アメリカ」に加えて重要であると考えられるので優先度にしたがって重みを加えるようにする。

[参考文献]

[1] K. Cheng and Y. Kambayashi. Enhanced Proxy Caching with Content Management, Knowledge and Information Systems (KAIS), An International Journal. April 2002, 4(2): 202-218

[2] K. Cheng, Y. Kambayashi, ``A Semantic Model for Hypertext Data Caching," 21st International Conference on Conceptual Modeling (ER2002), October 7-11, 2002, Tampere Finland. (to appear)

[3] R. Lee, H. Takakura, and Y. Kambayashi, "Visual QueryProcessing for GIS with Web Contents," Proc. of the 6th I FIP WorkingConference on Visual Database Systems, pp.171-185, May 29-31, 2002.