2J2-NFC2-4

情報可視化におけるグラフ緻密化のための Query Reformulation

Query Reformulation for Graph Densification on Information Visualization

堀部 高史[*1] 福本 淳一[*2]
Takafumi Horibe Junichi Fukumoto

[*1]立命館大学大学院理工学研究科情報理工学専攻
Graduate School of Science and Engineering Advanced Information Science and Engineering Major, Ritsumeikan University

[*2]立命館大学情報理工学部メディア情報学科
College of Information and Engineering Department of Media Technology, Ritsumeikan University

In our previous approach, we proposed a numeric information extraction method from Web documents for graph generation. However, the generated graph sometimes has sparse parts which have no data. It will be necessary to fill additional information to make the graph precise one. In this paper, we propose a query reformulation method for graph densification. We have conducted two kinds of experiments to reformulate queries for extraction of necessary information. We have successfully got additional information for graph densification and regenerated the graph using the new information.

1. はじめに

文書中から商品の価格や日付などの数値表現を抽出し, 表やグラフを用いて視覚的に表す技術として「動向情報の要約と可視化に関するワークショップ（MuST）」[松下 05] を中心に研究が行われている. MuST では, タグが付与されたコーパスを対象に様々な可視化が試みられている.

我々は, タグが付与されていない Web 文書から可視化に必要な情報を自動的に抽出し, それらの情報からグラフ作成などの可視化の研究を行ってきた [堀部 09]. Web 文書から抽出する情報としては, まず, 量や価格を表す 数値, その数値の種類を示す 属性, その数値の 対象 の 3 つがある. さらに, その情報の時間情報として 日付 も抽出する. これらの情報の抽出のため, 人手でパターンを作成し, 抽出された情報をグラフ化する手法の提案を行ってきた.

本手法により抽出された情報のグラフ化の際, プロットされた点に疎な部分が存在する場合がある. このようなグラフをより正確なものにするためには, グラフの疎な部分を補う情報が必要である. 本論文では, グラフの疎の部分の情報補完, 緻密化のためのクエリを自動生成する手法について述べる.

2. グラフの疎の情報補完手法の概要

情報補完のためのクエリ生成として, まずグラフの縦軸と横軸のどちらに注目して補完するかを判断する必要がある. 本手法では, 縦軸でのクエリ生成は, 数値の範囲を限定できないため, 日付の範囲が定まっている横軸でのクエリ生成を行うものとする.

グラフ中の疎な日付間隔を補うためのクエリ生成は, プロットされた点間で相対的に疎な期間を対象にする. 本手法では, 疎な期間を 全ての日付間隔の平均の 2 倍以上 の部分のうち最も長い期間をクエリ生成の対象とする.

本研究では, 2 種類のクエリ生成手法を提案する.

- 手法 1: 中間点の日付でのクエリ生成

  疎の期間の中間点の日付でクエリを生成する. 図 1 に示すように, 疎の期間が「3 日〜30 日」の場合, 中間点は「16.5 日」となり, 切り上げた日付の「17 日」である.

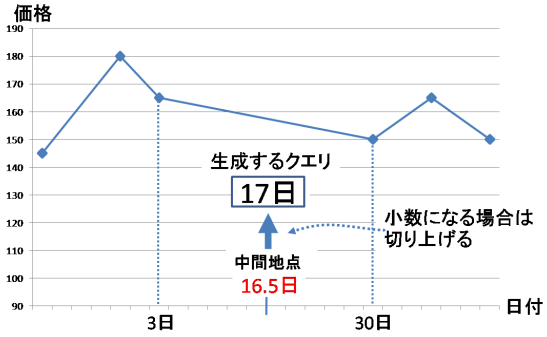

図 1: 中間点の日付でのクエリ生成の例

- 手法 2: 一定幅の日付間隔でのクエリ生成

  疎の期間の古い日付から一定の日付間隔で区切り, 各点でクエリを生成する. 日付間隔はグラフ中の最小の日付間隔とする. 図 2 に示すように, 最小の日付間隔は 5 日であり, 疎の期間で, 5 日ごとに 5 件のクエリが生成される.

検索のためのクエリ生成は, 決定された日付情報に対象, 属性を組み合わせたものを用いる. 生成されたクエリを用いた検索により得られた記事から可視化のための情報を抽出し, それらをグラフ作成のための情報に追加し, グラフの再描画を行う. この場合, 検索のために決定した日付以外の情報も抽出される場合があるが, それらもグラフ再描画に用いる. グラフの再描画後に以下のグラフの疎の条件を満たす部分がなくなるまで, クエリの生成を繰り返す.

- 全ての日付間隔で疎が解消された場合
- 再描画しても疎の期間の情報が更新できなかった場合 (ただし, 他に疎の期間が存在する場合は, 次に大きな疎の期間のクエリ生成を行う)
- 再描画後の疎の期間が, 最初のクエリによるグラフの平均日付間隔を下回った場合

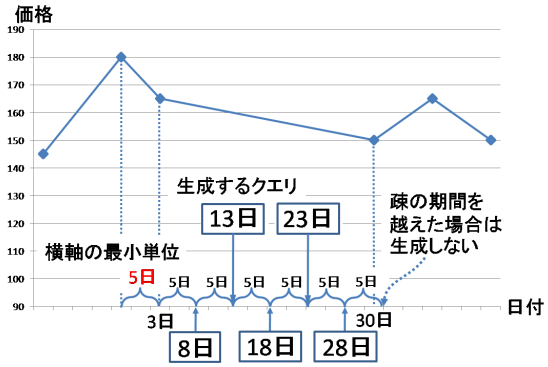

図 2: 一定幅の日付間隔でのクエリ生成の例

- 横軸の最小の日付間隔が「1 日」以下になった場合 (手法 2 の場合のみ)

3. 実験・評価

実験では, 記事検索に Google を使用し, 検索結果上位 25 記事を利用し, その記事からパターンによる可視化情報抽出を行う. 今回は図 3 の点線のグラフから, 本研究の提案手法のうち, 「中間点の日付でのクエリ生成手法」の有効性を検証する. 最初のクエリによる抽出結果の詳細を表 1 に示す.

表 1: 最初のクエリによる抽出情報

対象	属性	数値	日付
レギュラーガソリン	全国平均価格	155.5	20071210
レギュラーガソリン	全国平均価格	180	20080709
レギュラーガソリン	全国平均価格	139.5	20081022
レギュラーガソリン	全国平均店頭価格	151.3	20081027
レギュラーガソリン	全国平均価格	128.8	20090907
レギュラーガソリン	全国平均価格	123.5	20090909

表 1 より, 疎と判定された中で最も長い期間は, 「2008 年 10 月 27 日〜2009 年 9 月 7 日」である. この期間の中間点は「2009 年 4 月 2 日」で, 生成するクエリは「レギュラーガソリン 全国 店頭平均価格 2009 年 4 月 2 日」である. このクエリを Google に入力し, 検索結果の上位 25 記事を取得し, 可視化情報抽出を行う. 新たに追加された情報を表 2 に示す.

表 2: クエリ生成による新たに追加した抽出情報

対象	属性	数値	日付
レギュラーガソリン	全国平均価格	142.2	20080401
レギュラーガソリン	全国平均価格	108.3	20090202
レギュラーガソリン	全国平均価格	109.4	20090216
レギュラーガソリン	全国平均価格	109.6	20090223
レギュラーガソリン	全国平均価格	111.5	20090330
レギュラーガソリン	全国平均価格	113.2	20090406
レギュラーガソリン	全国平均価格	115.1	20090420
レギュラーガソリン	全国平均価格	115.9	20090427

新たに追加した情報に, 最初のクエリによる抽出結果と同一の情報が含まれる場合は, 除く. また, クエリ生成で用いた日付, 疎の期間以外の情報が抽出された場合も, 新たな情報としてグラフ生成に用いることで, グラフを緻密化することができた. グラフ生成の結果を図 3 の丸と実線で示す.
グラフ再描画後, 終了条件の 3 番目に当てはまるため, 疎がないと判断され, クエリ生成を終了する.

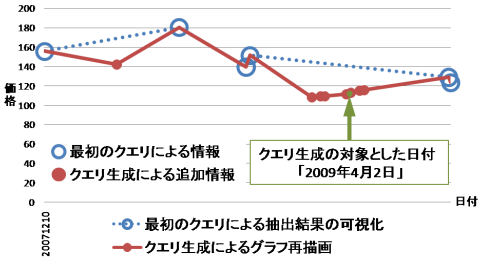

図 3: 抽出情報の可視化

4. 考察

中間点の日付で生成したクエリで新たに獲得した情報には, クエリで用いた日付が抽出できない場合や疎の期間以外の情報が抽出されることがある. その日付の記事が検索されなかったためであるが, 他の検索結果より, 疎の部分を補えるため, 問題はなかった.

クエリ生成の手法 1 と手法 2 では, 取得する記事数と記事内容の種類が異なる. これは, 生成するクエリの数によるためである. クエリの生成を繰り返すことにより, 最終的には差はないと思われ, 疎の情報補完という観点からは, どちらの手法が有効であるかは, これまでの実験からは, 判断できなかった.

新たに生成したクエリで再検索した結果, 同一日で異なった数値情報が抽出されることがある. これは, 日付情報の抽出ミスとその数値情報の情報源の違いによって起こることがほとんどであった. 今回の実験結果にはなかったが, 「来週には 130 円を超えるだろう」のような未来の表現に対する扱いも必要である.

5. おわりに

本研究では, グラフの「疎」の情報補完, 緻密化するために, 再検索のための 2 つのクエリ生成手法を提案した. その結果, 図 3 の実線に示すように「疎」となっていた期間は解消され, 有効性を示すことができた. しかし, 均等にプロットされている場合, 疎と判定されないため, クエリの生成が行われない. また, 今回は, 日レベルで推移するトピックでしか実験を行っていない. 今後は, 月レベル, 年レベルで推移するトピックに対する実験を行っていく予定である.

参考文献

[松下 05] 松下, 加藤: “動向情報に基づく情報可視化の基礎検討”, 2005 年度人工知能学会全国大会 (第 19 回) 論文集, 1E3-03. 2005.

[NExT 04] 渡辺, 桝井, 福本: “固有表現抽出ツール NExT の精緻化とユーザビリティの向上”, 言語処理学会第 10 回年次大会発表論文集, pp.413-415, 2004.

[堀部 09] 堀部, 福本: “情報可視化のための Web からの数値関連情報抽出手法”, 電子情報通信学会, 第二種研究会資料, IEICE SIG Notes, WI2-2009-55, pp39-40, 2009.10