平成２３年　１月３１日

専攻名	経営情報システム工学	学籍番号	07336887
申 請 者 氏 名	坪川貴和		
指導教員氏名	五島洋行		

審査委員主査　五島洋行

審査委員　　大里有生

審査委員　　中村和男

審査委員

審査委員

専攻主任印

論　文　要　旨

論文題目	Web ニュース上のホットトピックスの効率的な検索手法

本論文では，Web 上に膨大な数存在するニュース記事の中から，ユーザーの興味がある話題に関連する記事を効率的に検索する手法を提案する．Web ニュースから目的の記事を探し出す一般的な方法として，検索エンジンを用いた検索が挙げられる．検索エンジンを用いた Web ニュースの検索において，情報検索に不慣れな，または興味のある話題の動向に詳しくないユーザーが直感的に入力したキーワードでは，意図した内容の記事のみを素早く得ることは難しい．したがって，このようなユーザーであっても，効率的に目的の記事が得られる手法が必要とされている．そこで提案手法では，１．注目されている話題と，その注目されている期間に関するユーザーの理解を補助するグラフ，２．検索の際に用いることで，それらの話題それぞれを容易に取得可能にする単語，の二つをユーザーに提示する．

提示されたグラフをユーザーが参照した時に，特定の話題に興味を持ったならば，提案手法が提示した単語を検索エンジンに入力して検索を行うことで，その話題に関連する記事を効率的に検索することができる．つまり，情報検索に不慣れな，または話題の動向に詳しくないユーザーであっても，提案手法を用いることで，気になる話題に関連した記事を，容易に取得することができる．上記の１．と２．を生成するために，提案手法では記事集合に対していくつかの処理を行う．

まず，対象となる記事の集合を，内容が類似した記事同士で構成されるいくつかのグループに分類し，それらを各々のトピックスとして扱う．次に，ニュースの出現頻度のバースト性に注目し，それぞれのトピックスの注目度とその注目期間を算出し，それらの値を基にホットトピックスを選定する．選定されたホットトピックスより，上記 1. が生成される．最後に，ホットトピックスに関する記事の中から，それらの記事の特徴をよく表す単語を抽出する．抽出された単語より，上記 2. が生成される．また，抽出された単語を評価する評価関数を定義し，優先順位をつける．その評価関数では，注目度の高いホットトピックスを優先したり，時間の経過により情報の価値が減衰したりさせている．そして，優先順位の高い単語を，ユーザーに優先的に提案する．

提案手法が提示する単語を実際の検索エンジンに入力して検索する実験を行い，ホットトピックスに関連する記事が効率的に取得できることを確認した．

An Efficient Search Method for Obtaining Bursty Topics from Web News

Takakazu Tsubokawa
Faculty of Management and Information Systems Engineering, Nagaoka University of Technology Nagaoka

Abstract

In this thesis, we propose an efficient method for easily retrieving intended articles from a huge number of web news distributed over the internet. Here, the intended articles mean the ones that relates to a topic interested by users. A common method for obtaining intended articles distributed over the internet is to use a search engine.

Hence, we focus on efficiently retrieving web news using a search engine. In particular, we focus on users who are unfamiliar to the information retrieval or to the target topic. Then, we consider keywords intuitively determined by users who are unfamiliar to the information a case where such users determine keywords intuitively. If such keywords are inputted into a search engine, the users may not be able to easily obtain articles that relate to the target topic. Thus, irrespective of the expertise on information retrieval or familiarity with the target topics, it is required to develop a method for finding keywords by which the target articles can be obtained easily.

In view of this need, we proposed a method which provides the following two outputs: 1) an assistive graph for browsing bursty topics, points and time periods of the bursty topics, 2) assistive words by which the user can easily obtain articles that relate to the bursty topics.

We assume that a user is now interested in a specific topic by looking the assistive graph. For such case, the proposed method provides assistive words to the user. If these words are used for the retrieval, the user can easily obtain articles that relate to the interested topic.

For making the above two outputs, the proposed method performs several processes for the original articles. These are consisted of the following three steps. First, the proposed method classifies the articles into several groups based on the similarity of the contents. Each classified group is dealt as a topic. Second, the method focuses on the

bursty of the news as a document stream. Then, the bursty level and time period are calculated for the respective topics. Furthermore, bursty topics are selected from the classified topics considering the bursty levels. The assistive graph is generated based on the bursty level and the time period of the selected topics. In the last step, the method extracts words which describe the features of the articles related to the bursty topics. In addition, priorities are attached for the respective assistive words by defining an evaluation function. The proposed method preferentially gives the assistive words with higher priority to the user.

We confirmed the effectiveness of the proposed method through several experiments. The outputs of the proposed method are utilized for determining keywords which are inputted into a search engine. As a result, we can confirm the users can obtain target articles easily.