学位論文 博士（工学）

膨大な文書を対象とした
情報集約データベースに関する研究

2012年3月

慶應義塾大学大学院理工学研究科

富田準二

主　論　文　要　旨

報告番号	(甲)乙　第　　号	氏　名	富田 準二

主 論 文 題 目 :

膨大な文書を対象とした情報集約データベースに関する研究

（内容の要旨）

Web や検索エンジンの進歩によって誰でもが簡単に, 所望のページを取得できるようになってきている．しかしながら，例えば，会社や製品の評判や競合他社の動向などは，単一の文書としてまとめて記述されてはいないため，複数の文書を集め，その内容をまとめる作業（情報集約）を行わなければならない．現状，このような情報集約を行うための汎用的な枠組みは確立されていないため，情報集約サービスを実現するためには，個別のアプリケーションプログラムを最初から開発する必要がある．

本研究では，情報集約タスクを実行するための汎用的な枠組みとして，情報集約データベース（IADB: Information Aggregation DataBase）を提案する．IADB では，集約の対象となる情報の断片を，対象物とそれに付随する属性の集合（情報要素タプル）で表現する．例えば，評判情報であれば，“製品 A の画面は美しい．”という情報の断片を，<製品 A，画面，美しい，好評>（<対象物，評価属性，評価表現，評価極性>）という情報要素タプルで表現する．IADB は，このような情報要素タプルからなる仮想的な情報要素リレーションを大規模な文書から自動的に生成し，そこへの検索と集計を行うことで，様々な情報集約タスクを実行できるようにする．本研究では，特に，IADB を構築するうえでの技術課題として，(a)文書から対象物を抽出するための辞書の自動構築手法，(b)事前に抽出した情報要素の属性と，対象物を表す入力キーワードとを用いた情報要素リレーションの動的生成手法，(c)情報集約に特化した独自の問合せ言語，のそれぞれについて検討し，設計・実現する．

IADB を評判情報の集約を行う実サービスに適用し，情報要素リレーションへの簡易な問合せによって，様々な有用な情報集約結果を取得できることを示す．また，各技術課題に対して提案手法は，(a)に関しては，特に多義語の語義の網羅的な収集に有効であること，(b)に関しては，入力キーワードが未知語であったとしても，実時間で情報要素リレーションを生成できること，(c)に関しては，表記ゆれなどに対応しながら階層的な内訳をもつ集約結果を簡易な記述で取得できることを示す．更に，IADB が，他の情報集約タスクにも適用できる汎用的な枠組みであることを述べる．このように，IADB を用いることで，新商品の評判のような今まですぐには取得できなかった情報を多面的かつ即座に提供するオンラインサービスを，少ないコストで実現できる．

SUMMARY OF Ph.D. DISSERTATION

School Keio university	Student Identification Number	SURNAME, First name TOMITA, Junji

Title

A Study on an Information Aggregation Database for a Large Number of Documents

Abstract

Web and search engines enable us to obtain required documents easily. However, aggregating the fragments of information in a large number of documents is needed for obtaining the information that is not written in a single document such as the repetition of a company or a product, and the strategy of competitors. Since there is no concrete framework for aggregating the fragments of information, developers have to implement a specific program for such an aggregation task.

This research proposes an information aggregation database (IADB) for executing such tasks. In the IADB, an information fragment can be represented as a tuple consisting of a target object and its attributes. For example, for sentiment information, a fragment "The display of Product A is clear." is represented as a tuple "<Product A, display, clear, positive> (<target object, sentiment property, sentiment expression, sentiment orientation>)". The IADB enables to execute such aggregation tasks by automatically generating a virtual relation of the tuples from a large number of documents, and searching and summarizing the relation. This research deals with three technical issues; (a) automatic construction of a dictionary for extracting target objects in documents, (b) online relation generation based on combining pre-extracted attributes and a given keyword, and (c) a query language especially designed for such tasks.

The trial of applying the IADB to a real sentiment analysis service shows that simple queries to the relation can generate effective aggregation results for the service. Evaluations also show (a) is effective for extracting all the meanings of multi-meaning words, (b) allows realtime generation of the relation for an unknown word, and (c) makes a hierarchical result having sub-total of aggregation and handles the same word in different forms easily. Furthermore, this research also shows the IADB is general enough to apply to other tasks. Therefore, the IADB enables developers to realize online services with small amount of effort, which produce multiple views of aggregated information such as the repetition of a new product.