ARTICLE IN PRESS

Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Automatic faceted navigation

Bei Xu, Hai Zhuge *

Institute of Cyber–Physical–Social Intelligence, Nanjing University of Posts and Telecommunications, 210046, China
Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190, China

ARTICLE INFO

Article history:
Received 5 September 2012
Received in revised form
10 November 2012
Accepted 7 December 2012
Available online xxxx

Keywords:
Semantic unit
Human reading process
Faceted navigation
Text analysis
Web search

ABSTRACT

Automatic faceted navigation provides a set of navigation operations for users to browse one text or more texts from different facets of content without any input. One sentence or a sequence of sentences on a relatively independent and complete content in text constitutes a semantic unit. Semantic units on one topic may be in different locations and there are distances between semantic units. A semantic facet is a set of semantically close semantic units representing one aspect meaning of a text. This paper proposes a faceted navigation mechanism that can extract semantic units from texts, build a hierarchy of semantic facets, and navigate the hierarchy, so that users can obtain interested contents from text and navigate among semantic facets without any prior input. The proposed mechanism goes beyond cyberspace by considering the human reading processes. Experiments show the effectiveness of the proposed mechanism.

1. Introduction

Efficiently obtaining needed content from the ocean of text in cyberspace is a challenge.

One method is to create a full text summarization [1]. However, it is hard for these mechanisms to provide multi-facet content. The summarized content may not be significant for users. Multiple-document summarization [2,3] extracts content from multiple texts on a common topic. Since the topics of multi-document summarization are preset and prominent, the summarization cannot cover non-prominent content which may be needed by users. The problem still remains in topic extraction methods [4,5] which are to extract prominent topics within a text. Information retrieval systems on text, like digital libraries, can gather and organize multiple article content to provide well-defined services for users [6–8]. But most of the systems fail to satisfy users' multiple faceted needs [9].

Some researches focus on facilitating users' behaviours for information retrieval. The first kind requires users to express their preferences through profiles or settings [6–8]. Set-ups and maintenance of the profiles or settings limit users' reading efficiency. The second kind requires users to input several keywords to represent their interests on one aspect of content [10–12]. However, a user can hardly input appropriate keywords when he/she has little prior knowledge on a text. Collaborative tagging allows users to freely tag resources for sharing and exploring content [13]. However tags are hard to represent content, especially multi-facet content. The third are the methods which can directly analyse text without prior input. These kinds of methods need no pre-conditions. However, these methods do not provide any faceted navigation mechanism for users and lack interactive operations.

The idea of faceted navigation is proposed to satisfy users' faceted needs on text. Faceted navigation can provide multiple aspects of content in text and convenient operations. Faceted navigation on text can improve users' reading efficiency and cover the texts' full content. However, previous methods on faceted navigation were not suitable for dealing with plain text. These methods mostly deal with structural or semi-structural data, such as text-annotated data [14], RDF data [15] or images [16]. As for text, these faceted navigation mechanisms are basically based on analysis of a text's attributes (e.g., an article's author, publishing time, volume, number, and number of pages, etc.) while neglecting the human reading process.

From the point of view of methodology, previous techniques on text can be divided into three kinds. The first is based on statistic analysing methods of text, such as word frequency [17], position of words in text [18], etc. Some summarization approaches represent text as a network of language components (words, sentences, and paragraphs) and use strategies based on different network features, such as degree, short path [19], *k*-cores [20], etc. The second is based on training and testing data, including vector space models [21], probabilistic models [22], latent semantic analysis [23], etc. The third is based on background knowledge, such as ontology-based methods [24], encyclopaedic knowledge Wikipedia [25], etc. The mentioned techniques endeavour to dig out static or background features from text while neglecting cognition factors. The

* Correspondence to: No. 6, Science South Rd., Zhong Guan Cun, Haidian, Beijing, 100190, China.
E-mail address: zhuge@ict.ac.cn (H. Zhuge).

0167-739X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2012.12.003

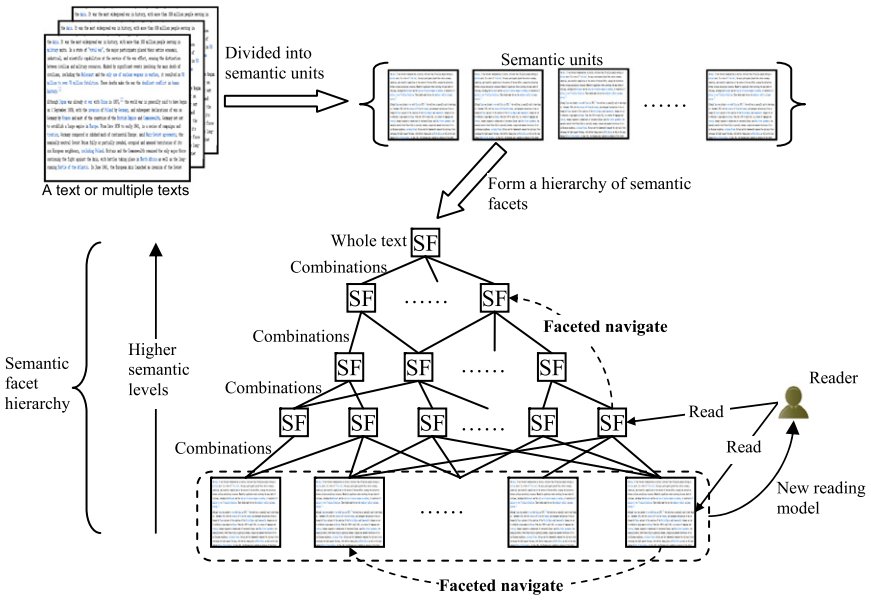

Fig. 1. The mainframe of faceted navigation without prior input. *SF* denotes a semantic facet.

importance of interaction in forming social semantics is pointed out in [26].

Cognitive scientists are concerned with the human cognitive process and regard the process of constructing semantics of text as the process of constructing mental spaces [27,28]. However, they have not proposed applicable approaches. The insight of a semantic lens is proposed to support individuals to view content from multiple dimensions and different abstraction levels and freely jump between dimensions. The idea of the semantic lens on texts is discussed in [26,29].

We have proposed a faceted navigation approach through keyword interaction in [10]. Users can input keywords reflecting their preferences and then obtain a semantic unit from text consisting of sentences semantically close to the keywords. The sentences in a semantic unit are discrete.

This paper proposes a faceted navigation mechanism incorporating the simulation of the human reading process. The mechanism enables users to read multiple facets of one text or more texts without any prior input. One facet displays one aspect of content of the text. The mechanism has the following features:

(1) Users can simply and directly get faceted content without any user profile setting or inputting any keyword.
(2) A new kind of semantic unit consisting of consecutive sentences is proposed to represent relatively independent and complete topics in text.
(3) A hierarchy of the semantic units is built to provide a faceted navigation service for users.
(4) It can be used for multiple texts.
(5) It provides a new type of reading model for users.

2. Semantic unit and main frame of faceted navigation on text

A long article usually consists of content on several local topics. The topics shift from one to the other while reading and accompany many different emergence of words. The texts on a common topic should share some words.

Therefore an article can be regarded as sequential units of content with a topic shift. Each unit corresponds to a local topic. Semantic units have the following characteristics:

1. the sentences in a semantic unit are consecutive;
2. semantic units on different topics accompany many different words;
3. adjacent semantic units represent relatively independent and complete topics;
4. two non-adjacent semantic units may be on one topic.

A semantic unit from the ith sentence to the jth sentence in text is represented as $S(i, j)$. The semantic units' comprehensibility is guaranteed because of the first and second characteristics. The size of a semantic unit depends on the specific content of text.

The proposed mechanism can automatically search semantic units in an article based on simulation of the human reading process. The impressions of words are used to search semantic units. A reader can clearly remember words when the words are frequently mentioned and will forget words when the words have not been mentioned for a while. So the impressions of words change while reading. Semantic units can semantically merge into semantic facets which are combinations of semantic units. Semantic facets have different semantic levels. The higher a semantic facet's semantic level is, the more semantic units it covers.

We give the main frame of faceted navigation as shown in Fig. 1. Users can freely view the found semantic units' or semantic facets' content and faceted navigate between them.

3. Searching semantic units through simulation of the human reading process

3.1. Simulation of the human reading process

This section briefly introduces the human reading process and relevant concepts [10].

An article A can be represented as a set of ordered sentences: $A = \{s_1, s_2, \ldots, s_n\}$, where s_i denotes the ith sentence, and n denotes the total number of sentences in A. A sentence s is a set of words with no order represented as: $s = \{w_1, w_2, \ldots, w_k\}$, where w_i denotes a word, and k denotes the number of words in the sentence. The human reading process performs on a dynamic word network reflecting the circumstance of the reading sentence.

A *word network* of a text is a dynamic network denoted as $WN(x, D) = \langle Nodes, Links, Weight\_nodes, Weight\_links \rangle$, where the nodes in *Nodes* are a set of words in the text, *Links* is a set of links between nodes, *Weight_nodes* is a vector that records the weights of all nodes, and *Weight_links* is a vector that records the weights of all links. $D \in \{1, \ldots, n\}$. x denotes the network based on the xth