Data & Knowledge Engineering 100 (2015) 191–210

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

A novel methodology for retrieving infographics utilizing structure and message content

Zhuo Li [a], Sandra Carberry [a], Hui Fang [b], Kathleen F. McCoy [a], Kelly Peterson [a], Matthew Stagitis [a]

[a] Department Computer and Information Science, University of Delaware, United States
[b] Department of Electrical and Computer Engineering, University of Delaware, United States

ARTICLE INFO

Article history:
Received 10 January 2015
Accepted 28 May 2015
Available online 25 June 2015

Keywords:
Semi-structured data and XML
Information retrieval
Digital libraries
Query
Graphic retrieval
Natural language query processing
Short document expansion
Linear combination ranking model

ABSTRACT

Information graphics (infographics) in popular media are highly structured knowledge representations that are generally designed to convey an intended message. This paper presents a novel methodology for retrieving infographics from a digital library that takes into account a graphic's structural and message content. The retrieval methodology can be summarized thus: 1) hypothesize requisite structural and message content from a natural language query, 2) measure the relevance of each candidate infographic to the requisite structural and message content hypothesized from the user query, and 3) integrate these relevance measurements via a linear combination model in order to produce a ranked list of infographics in response to the user query. The methodology has been implemented and evaluated, and it significantly outperforms a baseline method that treats queries and graphics as bags of words.

1. Introduction

Information graphics (infographics) commonly appearing in popular media, such as bar charts and line graphs, are effective visual representations of a relationship between data entities. Designers of such graphics generally construct them using well-known communicative signals (e.g., coloring a bar differently to highlight it) to convey a high-level *intended message*. For example, the bar chart in Fig. 1 ostensibly conveys the message that Toyota has the highest profit compared to the other car manufacturers listed. Although it is possible to describe the content of an infographic by paragraphs of written explanations, it is easier for a reader to absorb the information quickly from a graphic [1], making infographics an important and unique knowledge source that should be accessible and retrievable based on their content. As a take-off from the proverb "a picture is worth a thousand words", we can similarly say that "a graphic is worth a thousand words" since it contains a multitude of information.

Yet compared to the retrieval of text documents [2,3] and pictorial images [4,5], scant attention has been given to the retrieval of infographics. Conventional search engines rely on the document text that contains the infographic, including the infographic's file name, the image tag from the webpage html source file, and words in the accompanying article appearing near the infographic in the source file. These approaches ignore the content of the infographic itself.

Consider the query, *"How does the net profit of major car manufacturing companies compare?"* This query is requesting infographics that convey a comparison of car manufacturing companies according to their net profit, as suggested in part by the use of the verb *"compare"* and the plural form of *"companies"* in the query. When this full sentence query was entered into Google Image Search on December 10th, 2014, no satisfactory graphics appeared among the top 10 infographics returned. The infographic deemed most

E-mail addresses: ivanka@udel.edu (Z. Li), carberry@udel.edu (S. Carberry), hui@udel.edu (H. Fang), mccoy@udel.edu (K.F. McCoy), keldryc@udel.edu (K. Peterson), mattstag@UDel.Edu (M. Stagitis).

http://dx.doi.org/10.1016/j.datak.2015.05.005
0169-023X/© 2015 Published by Elsevier B.V.

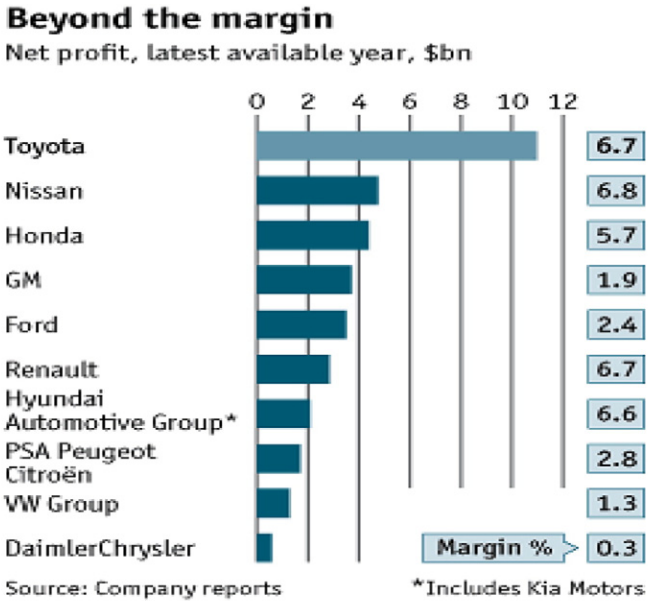

Fig. 1. Infographic example.

relevant was the bar chart shown in Fig. 2, which displays a car model (Volkswagon Golf) and its sales trend, not a comparison of car companies; the terms *"profit"*, *"car"*, and *"company"* appeared near the infographic in the accompanying text article and may account for its retrieval. On the other hand, a much more relevant graph is the one shown in Fig. 1; it presents a comparison of twelve car manufacturing companies by representing each company by a bar on the independent axis and ordering the bars according to the

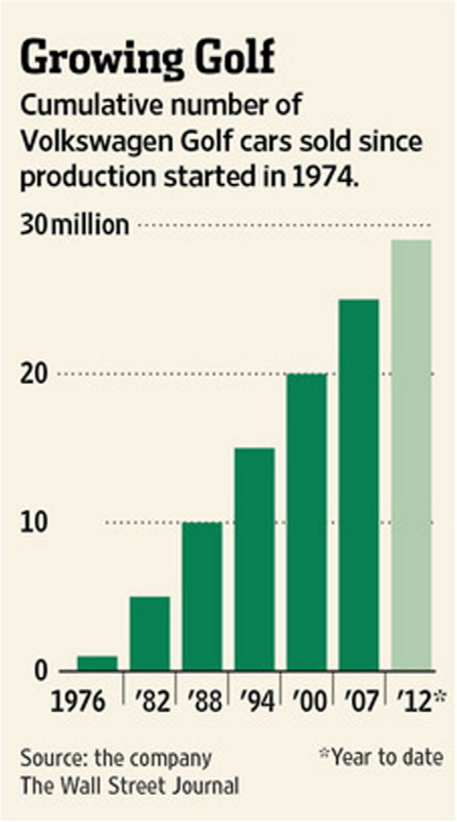

Fig. 2. Top retrieved infographic by Google Image.