

**Adaptable Extraction of Key Elements  
from Weblogs**

**Noriko Imafuji Yasui, Xavier Llorà,  
David E. Goldberg  
IlliGAL Report No. 2006024**

Illinois Genetic Algorithms Laboratory  
University of Illinois at Urbana-Champaign  
117 Transportation Building  
104 S. Mathews Avenue Urbana, IL 61801  
Office: (217) 333-2346  
Fax: (217) 244-5705

# Adaptable Extraction of Key Elements from Weblogs

Noriko Imafuji Yasui<sup>1</sup>   Xavier Llorà<sup>2</sup>   David E. Goldberg<sup>1</sup>

<sup>1</sup>Industrial and Enterprise System Engineering   <sup>2</sup>National Center for Supercomputing Applications  
University of Illinois at Urbana-Champaign   University of Illinois at Urbana-Champaign  
104 S. Mathews Ave. Urbana, IL 61801   1205 W. Clark Street, Urbana, IL 61801  
{nyasui,deg}@uiuc.edu   xllora@uiuc.edu

December 13, 2006

## Abstract

This paper proposes *AKEE* (Adaptable Key Elements Extraction) algorithm for web-log (blog, for short) mining. *AKEE* enables us to identify significant information in various blog components (e.g., a blog post, a series of blog posts, a set of series of blog posts, etc.).

## 1 Introduction

Blogs have changed the way people and organizations express, interact, and—quite unforeseen—exercise influence. On the context of market research, blogs are rich repositories of useful information—people’s firsthand voice. However, it would take more than a lifetime to read all the available blogs. The methodological approaches for efficiently finding relevant information from blogs are strongly demanded. Since a first large scale analysis of blog space by (Kumar, Novak, Raghavan, & Tomkins, 2003) was appeared, blog related research field has been quickly growing. Developing a system for modeling, measuring and analyzing blogs for market analysis support, we have also started to expand our on-going research project DISCUS (Goldberg, Welge, & Llorà, 2003) into the field.

In this paper, we propose *AKEE* (Adaptable Key Elements Extraction) algorithm for simultaneously finding two types of key elements (*container* and *content*) from a focused blog *component* (see Figure 1). For example, *AKEE* algorithm finds key sentences—*containers* and key terms—*contents* in a focused blog post—*component*. One of the biggest advantages of *AKEE* is its adaptability. *AKEE* is “adaptable” in the sense that we can use *AKEE* for various triplets of (component, container, content), for example, (a set of blog posts, blog post, sentence), (a blog post, paragraph, sentence), and so on. *AKEE* identifies significant parts in targeted information source.

## 2 *AKEE*: Adaptable Key Elements Extraction

In this section, we propose *AKEE* (Adaptable Key Elements Extraction) algorithm. This algorithm is based on a weighted directed bipartite graph induced from component-container-content relation (see Figure 1). Its edge weight assignment is also proposed here.

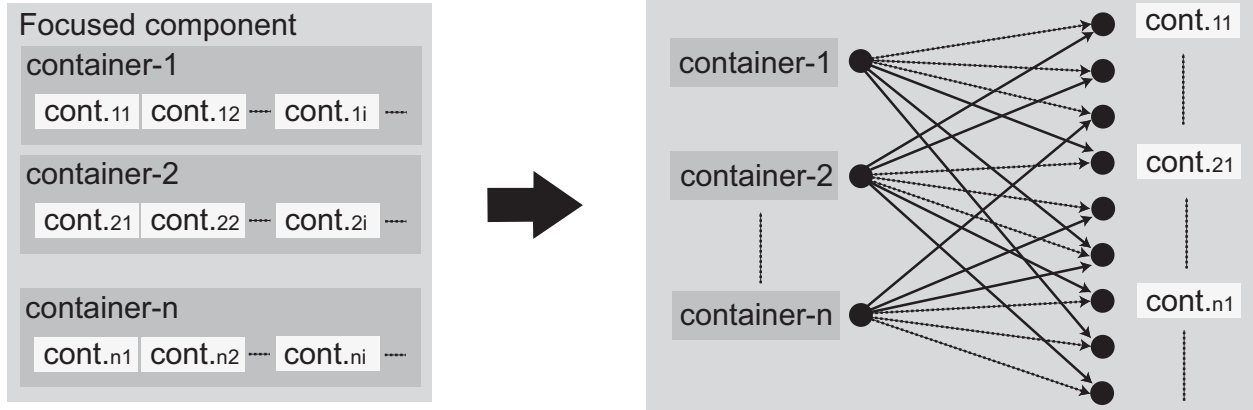


Figure 1: Component-container-content and its induced directed bipartite graph.

## 2.1 AKEE algorithm

AKEE finds *key containers* and *key contents* by scoring the containers and the contents in the context of their *significances* in the focused component. Higher scored containers and contents are *key containers* and *key contents* in the focused component.

AKEE uses HITS (Hyperlink-Induced Topic Search) algorithm (Kleinberg, 1999) in an unintended way. HITS is an algorithm for ranking web pages in contents of *hubs* and *authorities*. A good *hub* is a page that points to many good authority pages, and conversely, a good *authority* is a page that is pointed to by many good hub pages. AKEE is an algorithm for applying HITS framework to text mining. AKEE is based on mutually reinforcing relationship between containers and contents: significant containers in a focused component are the containers containing many significant contents, and conversely, significant contents are the contents contained in many significant containers.

A component is represented by a weighted directed bipartite graph  $G(V, E)$  where  $V$  and  $E$  are sets of nodes and weighted edges, respectively. Let  $V_H$  be a set of containers in a component, and  $V_A$  be a set of contents contained in the components.  $V = V_H \cup V_A$ ,  $V_H \cap V_A = \phi$ . Let denote an edge between  $h_i \in V_H$  and  $a_j \in V_A$  and its weight by  $(h_i, a_j)$  and  $w(h_i, a_j)$ , respectively. We define two types of edges; *direct-* and *indirect-*relation induced edges. The former and the latter are depicted by dotted and solid lines in Figure 1, respectively. *The direct-relation induced edges* are edges from container nodes to the nodes of their contents. *The indirect-relation induced edges* are edges from container nodes to the content nodes of another containers.

The edge weight  $w(h_i, a_j)$  is defined by the following equation.

$$w(h_i, a_j) = \begin{cases} 1, & \text{if } (h_i, a_j) \in E_D \\ sim(h_i, a_j), & \text{if } (h_i, a_j) \in E_I \end{cases} \quad (1)$$

$E_D$  and  $E_I$  are a set of direct- and indirect-relation induced edges, respectively.  $sim(h_i, a_j)$  is given by a term-based similarity between  $h_i$  and  $a_j$ . Let sets of terms in  $h_i$  and  $a_j$  denote by  $T(h_i)$  and  $T(a_j)$ , respectively. A term-based similarity between  $h_i$  and  $a_j$  is defined by

$$sim(h_i, a_j) = \frac{|T(h_i) \cap T(a_j)|}{|T(a_j)|}. \quad (2)$$

Containers and contents are ranked by *key scores* of containers (or *container scores* for short) and *key scores* of contents (or *content scores* for short). Let  $s(h_i)$  and  $s(a_i)$  denote the key score of

container  $h_i$  and the key score of content  $a_i$ , respectively. Similarly to HITS algorithm (Kleinberg, 1999), the mutually reinforcing relationship in AKEE algorithm are as follows: If the container  $h_i$  contained many contents with high key scores, then it should receive a high container score; and if the content  $a_i$  had been used by many containers with high key score, then the content should receive a high content score.

AKEE algorithm obtains container and content scores simultaneously by an iterative calculation. Given container score  $s(h_i)$  and content score  $s(a_j)$ ,  $s(h_i)$  and  $s(a_j)$  are updated by the following calculations.

$$s(h_i) \leftarrow \sum_{(h_i, a_j) \in E} s(a_j) \cdot w(h_i, a_j) \quad (3)$$

$$s(a_i) \leftarrow \sum_{(h_i, a_j) \in E} s(h_i) \cdot w(h_i, a_j) \quad (4)$$

AKEE algorithm is as follows. A vector of container scores and a vector of content scores are represented by  $S_H$  and  $S_A$  respectively.  $k$  in the below is a natural number. Highest ranked elements in  $S_H^k$  and  $S_A^k$  are key containers and key contents in the focused component.

**AKEE algorithm:**

1. Initialize  $S_H^0 = 1, 1, \dots, 1$ , and  $S_A^0 = 1, 1, \dots, 1$
2. For  $i = 1, 2, \dots, k$ 
  - (a)  $S_H^i$  is obtained using Equation (3) with  $S_A^{i-1}$
  - (b) Normalize  $S_H^i$  so the square sum in  $S_H^i$  to 1
  - (c)  $S_A^i$  is obtained using Equation (4) with  $S_H^i$
  - (d) Normalize  $S_A^i$  so the square sum in  $S_A^i$  to 1
3. Return  $S_H^k$  and  $S_A^k$

Kleinberg proved theorems that  $S_H$  and  $S_A$  converge and the limits of  $S_H^k$  and  $S_A^k$  are obtained by the principal eigenvectors of  $A^T A$  and  $AA^T$  (Kleinberg, 1999).  $A$  is an adjacency matrix;  $(i, j)$  entry is 1 if  $(h_i, a_j) \in E$ , and is 0 otherwise. Empirically,  $S_H$  and  $S_A$  converge very rapidly.

### 3 Experiments Overview

As experimental evaluation of AKEE, we analyzed the blog posts published on the Google Blog<sup>1</sup> from November 10th until December 7th. We examined AKEE with triplets (component, container, content) = (a blog post, sentence, term) and (a series of blog posts, a blog post, sentence). The former triple finds significant sentences and terms in a blog post, and the latter finds significant blog posts and sentences in a series of blog posts. Based on AKEE with the triplet (a series of blog posts, a blog post, sentence), we also performed time series analysis. Space in this paper did not permit us to insert a detailed description of the methodology and results.

We just briefly show one of the results using the triplet (a blog post, sentence, term). AKEE induced key sentences from a post entitled “*Old world meets new on Google Earth*”<sup>2</sup> were

<sup>1</sup><http://googleblog.blogspot.com/>

<sup>2</sup><http://googleblog.blogspot.com/2006/11/old-world-meets-new-on-google-earth.html>

*“I was able to explore and fly around the old maps and use the transparency slider to compare the old world and the new; as I did this, I thought to myself that this is the perfect marriage of historic cartographic masterpieces with the innovative contemporary software tools of Google.”*

and its highest ranked key terms were *map, earth, historic, explore, old, world, tool, cartography*. Reading through the post, the key sentences and terms obtained by AKEE were actually significant elements of the post.

## 4 Concluding Remarks

This paper proposed AKEE algorithm, which enabled us to identify significant information in various blog components (e.g., a blog post, a series of blog posts, a set of series of blog posts, etc.). In the poster, we will show various interesting results in detail.

Current edge weight assignment used in AKEE is quite simply using the term-based similarity. Our future work includes examining better edge weight assignments well-defined by a feature based-similarity. Furthermore, our approach to the analysis of the post is based on statistics instead of more traditional approaches based on natural language processing—from which we may benefit in future stages.

## Acknowledgments

We would like to thank to HakuHodo inc. for the project collaboration. This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant AF9550-06-1-0096 and AF9550-06-1-0370. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

## References

- Goldberg, D. E., Welge, M., & Llorà, X. (2003). *DISCUS: Distributed Innovation and Scalable Collaboration In Uncertain Settings* (IlliGAL Report No. 2003017). Urbana, IL: University of Illinois at Urbana-Champaign, IlliGAL.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604–632.
- Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). On the bursty evolution of blogspace. In *WWW '03: the 12th international conference on World Wide Web*