Chad Hogg
Review of "Web Mining Research: A Survey", by Kosala and Blockeel

This paper describes the entire spectrum of current (in 2000) research in areas that could be classified as "Web Mining". They subdivide this spectrum into three distinct categories of mining, based on the types of information that are used. These three categories are Web Content, which operates on textual information that is available through the Web; Web Structure, which operates on information about the way pages are organized into web sites and the links between them; and Web Usage, which operates on information about how users consume the contents of the web. Additionally, they divide content mining into a group of techniques that operate from an information retrieval standpoint and a second that use a database approach. In all cases, many examples are provided of recent research in the area, how the researchers represented data, and what techniques they employed.

As a source of references for someone interested in finding the particular research area of web mining that they would be interested in, this paper is a valuable resource. It is exhaustive, covering not only the important papers on traditional content and link analysis algorithms used by search engines, but a number of other topics in which similar techniques are employed on different types of data related to the web. In any particular subcategory of interest, there are listings of important research findings in the area and reasonably detailed information about the specific techniques used. Particularly, tables 3 and 4 provide a remarkably clear picture of how important research is progressing in two of the most proliferous research areas. These tables should be effective at helping readers combat the same issue of information overload addressed by the paper in the arena of research literature rather than web content.

As a source of technichal information about web mining methodologies, this paper is quite weak. The authors rarely describe a technique in enough detail for the reader to have any idea of its motivation or operation. Instead, they merely mention work that has been done, provide a reference, and move on to a different topic. For some publications this may be the desired state of a survey paper, but I suspect others will want a bit more information. A more focused and detailed paper, such as the one by Chakrabarti for the January 2000 edition of SIGKDD Explorations would be appropriate in those cases.

Additionally, this paper is clearly not ready for publication. A simple perusal of the abstract provides a glimpse into the grammar of the entire paper, with phrases such as "... there is a lot of confusions when comparing research ..." and "... point out some confusions regarded the usage of the term ...". Although few of the errors are more substantial than these, they make the paper quite difficult to read.

In summary, this paper is not ready for publication. However, a thorough proofreading by an expert in the English language and consideration to more explanatory content could make it a useful summary of applications of data mining techniques to various aspects of the web.