

The Design and Implementation of the Searching System of Integrated Digital Contents

陳韋任、邱紹豐

E-mail: 364924@mail.dyu.edu.tw

ABSTRACT

The application of digital-content files offers different function of key-word searching. But as the quantity and types of file increase, the application needs to be restarted several times when doing searching job. The most applications only offer simple key-word searching, and users can't add logic operators in searching strings. In our study, we implement a searching platform which combines with multiple modules of digital-content searching. Users can enter logic operators and searching strings by a universal interface to achieve multiple files and types of searching. To support the function of changing the weight of logic operators by adding parentheses in searching string, there will be a problem of dealing with key word sequence. Our approach is to transfer searching string to binary logical tree, then go through it by inorder traversal and postorder traversal to process logic operator and parentheses. At last, if the quantity and types of file are different, there will be several different format output results. Solving this problem, we design a wrapper which transfers all the results to XML format by standard Schema, and the application will read XML format and output it to user interface.

Keywords : keyword search、 search module

Table of Contents

封面內頁 簽名頁 中文摘要 iii ABSTRACT iv 誌謝 v 目錄 vi 圖目錄 viii 表目錄 x 第一章 序論 1 1.1研究背景與動機 1 1.2研究目的 1 1.3研究範圍 2 1.4論文各章題要 2 第二章 相關研究 3 2.1結構性檔案 3 2.1.1 BANKS 3 2.1.2 DBXplorer 5 2.2非結構性檔案 7 2.2.1 PDF檔案存取 8 2.2.2 Microsoft Office檔案存取 13 2.3半結構性檔案 20 2.3.1 結構式搜尋 20 2.3.2 關鍵字搜尋 23 第三章 研究方法 25 3.1介面模組 26 3.2樣板產生器 30 3.2.1 結構性檔案處理 30 3.2.2 非結構性檔案處理 32 3.3處理核心 35 3.4封裝器 35 3.5搜尋模組 38 3.5.1 微軟OFFICE系列檔案(Access除外) 38 3.5.2 Access檔案 39 3.5.3 PDF檔案 40 第四章 實驗結果 41 第五章 結論與未來發展 48 參考文獻 49

REFERENCES

- [1] Agrawal. S, Chaudhuri. S, Das. G, DBXplorer: " A system for keyword-based search over relational database " , Proceedings of the 18th International Conference on Data Engineering, San Jose, pp. 5-16, March, 2002.
- [2] Bhalotia. G, Hulgeri. A, Nakhe. C, Chakrabarti. S, Sudarshan. S, " Keyword searching and browsing in databases using BANKS " , Proceedings of the 18 th International Conference on Data Engineering, San Jose, pp. 431-440, March, 2002.
- [3] D. Florescu, I. Manolescu, " Integrating Keyword Search into XML Query Processing " , 9th WWW Conf., 2000.
- [4] Blakeley, J.A. " Universal data access with OLE DB " ,Compon '97. Proceedings, IEEE, pp.2-7, Feb, 1997.
- [5] Hassan, M.,Alhajj, R., Ridley, M.J., Barker, K, " Database selection and keyword search of structured databases: powerful search for naive users " ,Information Reuse and Integration, 2003. IRI 2003. IEEE International Conference ,pp.175-182,Oct. 2003 [6] Hristidis. V, Papakonstantinou. Y, DISCOVER " Keyword search in relational databases " , Proceedings of 28th International Conference on Very Large Data Bases, Hong Kong, pp 670-681, August, 2002.
- [7] Bruno Lowagie, " iText in Action, " Manning Publications; Second Edition edition,November 22, 2010.
- [8] Sahil Malik, " Pro Ado.net 2.0, " Apress,September 20, 2005 [9] Bruno Lowagie, itext, <http://itextpdf.com/> [10]IKVM, <http://www.ikvm.net/> [11]Microsoft, OLE Compound Document, [http://msdn.microsoft.com/en-us/library/windows/desktop/ms693383\(v=vs.85\).aspx](http://msdn.microsoft.com/en-us/library/windows/desktop/ms693383(v=vs.85).aspx) [12]Microsoft, ODBC 的基本概念(ODBC), [http://msdn.microsoft.com/zh-tw/library/thzsea08\(v=vs.90\).aspx](http://msdn.microsoft.com/zh-tw/library/thzsea08(v=vs.90).aspx) [13]Microsoft, Microsoft OLE DB (OLE DB), [http://msdn.microsoft.com/en-us/library/windows/desktop/ms722784\(v=vs.85\).aspx](http://msdn.microsoft.com/en-us/library/windows/desktop/ms722784(v=vs.85).aspx) [14]Microsoft, ADO.NET 概觀(ADO), [http://msdn.microsoft.com/zh-tw/library/h43ks021\(v=vs.80\).aspx](http://msdn.microsoft.com/zh-tw/library/h43ks021(v=vs.80).aspx) [15]Microsoft, Introducing the Office (2007) Open XML File Formats, [http://msdn.microsoft.com/en-us/library/aa338205\(v=office.12\).aspx](http://msdn.microsoft.com/en-us/library/aa338205(v=office.12).aspx) [16]Microsoft Office Word 2007, <http://office.microsoft.com/zh-tw/word-help/RZ010066490.aspx?section=29> [17]Apache, PDFBox, <http://pdfbox.apache.org/> [18]Apache, POI, <http://poi.apache.org/> [19]W3C, A Query Language for XML,(XML-QL), <http://www.w3.org/TR/NOTE-xml-ql/> [20]W3C, Extensible Markup Language (XML), <http://www.w3.org/XML> [21]W3C, An XML Query Language (XQuery), <http://www.w3.org/TR/xquery>

[22]W3C, XML Path Language (XPath), <http://www.w3.org/TR/xpath>