基於 Apache Spark 建構 XML Veracity 真實度之模型

中文摘要

近幾年大數據的數據量飛快地成長，超過 TB 等級的數據已是隨處可見，而數據傳輸在大數據中是一個重點探討的問題，大數據在做資料交會的時候，由於資料量龐大，所以資料的真實度不易掌控，也就是大數據 5V 當中的 Veracity 是我們最關切的問題。

而 XML(延伸標記語言 eXtensible Markup Language) 作為現今通用的網路資料交換格式，隨著網際網路資料的增長，也已經同樣具有大數據 (Big Data) 的特徵，在近幾年來，產業界與學界都將大數據列為重要研究議題，並投入相當多的資源支持大數據的研究。

本研究提出使用 Apache Spark 建構 XML Veracity 真實度模型以及真實度查詢模組，來解決大數據在資料傳輸當中我們所關心的真實度 Veracity 的計算問題，並可以讓使用者知道文件有多少可信度，XML 真實度模型計分是一個給定一個或多個 XML 文件，讓模型進行與標準 XML 文件做結構、版本或編碼的比較，並給訂一個分數的系統，本研究建構一個利用 Apache Spark 的平行化處理，加速真實度模型整體的運算速遞及效能的模組，讓使用者進行批次或是串流的 XML 文件上傳，再將上傳文件放入由 Apache Spark 建構的真實度計分模型來進行真實度計分，XML 文件真實度我們可以由幾個指標來判斷，例如文件結構、文件深度、文件內元素數量及標籤名稱相似度等，再使用 Apache Spark 所提供的 SparkSQL API，進行 XML 文件真實度查詢，查詢的功能可以從眾多文件中查詢與標準文件相似度最高的文件以及查詢有相關標籤名稱的文件，或是相同編碼的文件，抑或是在真實度積分當中同分或是相差多少分的文件，以上這一些功能的實作有助於使用者在面對大量 XML 資料的時候能夠有一個真實度計分的量化標準，在未來開發大數據應用的時候，可以對於在傳輸資料的時候有一個依據來知道資料的可靠程度。

Contents

List of Figures

1 introduction

1.1 背景

近年來數據以飛快的速度成長，TB 或是 PB 等級的數據隨處可見，在這資料快速產生且數據快速交換的時代，大數據一詞也越常被提及，國際數據公司 (International Data Corporation, IDC) 有研究指出，2008 年全球生產的資料量為 0.49ZB，2009 年全球產生的資料量為 0.8ZB，2010 年增長為 1.2ZB，2011 年的資料量更是已經達到 1.82ZB，這相當於全球每人生產 200GB 的資料，這麼龐大的資料也成了產業界與學術界所需要探討的重要議題，而有這麼大量的資料也意味著會有大量的應用會產生，而這一些應用當中一定會需要資料的交換，而在教會資料的時候，大多數的應用會選擇 XML。

在大數據中，有所謂的 5V，所謂的 5V 是指 Volume, Value, Veracity, Vleocity, Variety，Volume 是指產生的資料量