行政院國家科學委員會補助

大專學生研究計畫研究成果報告

計畫名稱：以語意分析及資料檢索技術為基礎之文件防抄襲系統研究

執行計畫學生：羅芳渝

學生計畫編號：NSC 101-2815-C-216-002-H
研究期間：101 年 07 月 01 日至 102 年 02 月 28 日止，計 8 個月

指導教授：應鳴雄

處理方式：本計畫涉及專利或其他智慧財產權，1 年後可公開查詢

執行單位：中華大學資訊管理學系

中華民國 102 年 04 月 24 日

以資訊檢索及語意分析為基礎的抄襲比對系統

摘要

近年來，網際網路及資訊科技的普及，人們皆能透過網際網路及資訊科技的幫助輕易取得各式各樣的資料、資訊、甚至知識。因而越來越多人取得資訊與知識的方式，逐漸由傳統圖書、文件等方式，轉變為透過資訊科技的方式。特別是透過資訊科技所下載取得的數位化資訊、資料，能輕易的經由複製、修改的程序，而改寫成自己的文件資料。也因為許多資料、資訊都能輕鬆透過網際網路及資訊科技取得並複製使用，因而延伸出越來越多的文件抄襲問題。

在各大專院中，每所學校的畢業要求門檻均不相同，就碩士和博士而言，大多必須完成學位論文方可畢業，對於資管系的大學生而言則需完成畢業專題才可畢業。近年來已發生了許多因論文抄襲被檢舉且查明屬實而追回碩士、博士學位的事件，因此在學生提出論文口試前，若能有一個有效的抄襲比對系統，將能減少這類問題的產生，而讓老師能專心於研究指導，不需花費精神與時間進行學生論文內容的抄襲比對。過去的研究偏重於內文文字的比較，缺乏對語意進行比較分析，因此無法針對程式設計這類課程的程式碼作業進行抄襲比對。然而，許多學生貪圖方便，經常將其他同學寫的內容，進行小幅度的修改，或是將程式碼的變數進行更換，導致傳統的防抄襲系統無法正確分析出這些抄襲的作業資料，也日益助長學生抄襲投機的心態。

網際網路及資訊科技的普及發達抄襲問題已日益嚴重，防範抄襲已成為當前重要議題，因此本研究嘗試以語意分析、資訊檢索為基礎，提出一個具備語意分析能力的文件防抄襲系統。最後再將本研究的雛型系統與市面上的論文防抄襲系統進行成效分析，並期望本研究之系統，能更正確的偵測出所有可能的抄襲文件，以嚇阻學生抄襲作業、論文之行為。

關鍵字：資訊檢索、語意分析、抄襲偵測系統

Development of a Plagiarism Detection System Based on Information Retrieval and Semantic Analysis

Abstract

With the rapid development of the Internet and information technology, the Internet and information technology more easily help humans to search data, information, or knowledge.

Thus, the more and more human obtain information and knowledge gradually from traditional library and documents instead of information technology. The master's and doctoral students, can easily to copy some words or sentences into their own document from someone’s document through Internet. The data and information can be easily obtained from Internet, so it was extending more and more plagiarism problem. The graduation requirements are difference among each school university, but the master’s and doctoral students general need to complete a thesis for graduation. In recent years, some master’s and doctoral theses was found have the plagiarism problem, therefore this study develops a plagiarism detection system based on information retrieval and semantic analysis to check the plagiarism document or paper out.

Keywords: information retrieval, semantic analysis, plagiarism detection system

一、 研究動機與研究問題

近年來，隨著資訊科技的進步，人們獲取資料或資訊的管道越來越多元且方便。在資訊科技未發達前，許多問題都需費時跑各大圖書館、書局以及學校查詢，而現今的資訊科技日新月異，想獲取資訊、解決問題，均可透過網際網路輕易取得答案，例如可透過線上圖書館、線上書局或是網際網路的搜尋引擎以獲取資訊，甚至可透過電腦將檔案下載儲存於電腦中，在資訊取得的成本越來越低廉時，抄襲、盜版、侵犯個人隱私等問題卻越來越嚴重。

本研究著重於作業與論文抄襲的問題，特別是科技便利的環境下，許多資料透過電腦便可輕鬆複製，日益嚴重的抄襲問題也嚴重影響了學生的學習成效，以及作業與論文的資訊品質。本研究希望能提出一個具備語意分析能力，並能以文件相似度的方式來判斷是否抄襲的防弊系統，以防止抄襲問題日益嚴重。現在市面上有少數網站提供論文抄襲的查詢系統，本研究將以幾個測試用的範本論文資料上傳至這些系統中，讓系統自動分析這些論文的抄襲程度，再與本研究之雛型系統進行比對，以確認本研究之成效。此外，現有的論文防抄襲系統，僅限於論文與網路上資料比較抄襲，並無法應用在教師平日教學的作業、報告的多檔案間的抄襲比對。因此，本研究嘗試以語意分析、資訊檢索等概念，提出一個抄襲比較系統。此系統不限於論文，亦可解決多位學生上傳作業、報告檔案的相互抄襲問題。

基於研究背景與動機，本研究目的歸納如下:

一、以語意分析及資訊檢索為基礎，設計一個抄襲偵測系統。

二、除了滿足目前網路上已有的論文抄襲檢測系統的功能外，亦增加多檔案相互抄襲的集群分析功能，將所有相互抄襲的文件，依據不同類型的相似性進行分類，並自動產生與其他文件相似的段落內容之檢測報告。

三、文件抄襲之比較分析，不局限於論文檔案之比較，本研究提出之方法，將可針對大專院校重點科目之作業進行檔案比較。

二、 文獻探討

一、 資訊檢索

所謂資訊檢索(information retrieval)是指從大量文件或是資訊中，搜尋並傳回使用者所想要的目標文件或是資料(Lancaster & Warner，1993)。資訊檢索是利用一些的設備和方法例如電腦科技，從龐大資料中查詢所需資訊的一種過程，其目的是利用電腦高度運算以及存取能力，來幫助使用者從大量缺乏結構化的資料中，快速取得所需資訊(鄭景俗等人，2008)。

資訊檢索最常使用的就是讓使用者輸入一個或多個關鍵字查詢的方式，將使用者所輸入的關鍵字跟資料中的文字進行比對，比對完後將最具有相關性或是完全吻合的資料傳回給使用者(鄭景俗等人，2008)。

二、 語意網

語意網是一種可提供電腦閱讀及理解網頁中涵義的一種網路內容形式，它有助於概念的溝通與知識體系的整合(黃居仁, 2003)。

目前的全球資訊網的是由統一資源識別碼（Universal Resource Identifier)、超文字傳輸協定(Hypertext Transform Protocol)以及超文字標記語言(Hypertext Markup Language）等要素所構成，人們必需透過這些資訊載體再經由超連結將資訊做串連，才可以看到想要之訊息，而這些資訊僅限於人類看得懂機器是無法知道涵義(應鳴雄、鄧光宏，2011)。

因此本研究必須透過語意網讓電腦閱讀並了解網頁中的涵義，再結合中文斷詞分析來增加比對的準確性。

三、 中文斷詞

中文文句相較於英文文句複雜許多，英文詞彙與詞彙間都有空格來做區隔，但中文詞彙卻是緊緊相連的(陳稼興等人，2000)。而對於一篇中文文章而言，如果要對文章內容進行分析，則必須先進行斷詞處理，再將內容中的詞彙進行分析與比對，才能真正了解文章中詞彙所要表達的意涵(應鳴雄、鄧光宏，2011)。

而為了能對於中文自然語言進行斷詞分析，中研院詞庫小組(CKIP) 建構了中文自然語言處理的資源與研究環境，為國內外中文自然語言處理及其相關研究提供基本的研究資料與知識架構(中央研究院資訊科學研究所，2004)。

本研究除了採用資料檢索方法進行相似文章的檢索分析外，也將對於文章中的每個句子進行結構和意義上的相互比較，因此必需先將文章內容透過 CKIP 系統進行斷詞分析，再將傳回來的詞彙資料進行分析與重組，以進行文章間的語意或概念抄襲比對。

四、 文件抄襲比對系統發展現況

目前較知名的文件抄襲檢查系統，主要包括香港中文大學 Hong Kong School Net(2012)的 VeriGuide 抄襲比對系統、雲書苑教育科技(2012)的在線論文抄襲比對系統(Plagiarism Verify System, PPVS)、PlagiarismDetect.com(2012)的 Detection Premium Plagiarism Detection System 等，主要是提供線上的文件抄襲比對服務，因此對於校園教學系統中，每次學生繳交的大量作業，卻較少針對群體作業間的文件抄襲進行比對分析。

因此，現有抄襲比對系統大多注重在上傳文件與網際網路上的資訊進行比較，對於文件與文件之間處理抄襲問題相當有限，大多使用關鍵字做為判斷依據。本研究希望能開發一套文件抄襲檢測系統，不僅能提供完整的文件抄襲檢測功能，亦能提供語意分析及高檢測率的服務。本研究除了能對於網際網路上的文件進行比較外，還著重在文件與文件之間的比較，並且不局限於特定領域的文件之中。

三、 研究方法

3.1 研究架構

本研究嘗試以語意分析、資訊檢索等概念，提出一個抄襲比較系統，以協助教師進行論文文件、學生作業的抄襲比對，減輕教師的負擔，而能將節省的時間運用在學生的論文與作業的指導上。本研究為了能使系統執行速度更有效率，將會先判斷文件與文件或是文件與網際網路上的資訊是否有相關，若有相關才進行抄襲的判定和剖析，因此本研究提出一個結合語意分析及資訊檢索技術的抄襲比對系統雛型架構，以便讓系統的抄襲辨識功能更為完善、精準。

3.2 系統架構

本研究為了能讓系統更有效率且精準的判定抄襲程度，因此採用語意分析來判定文章相關程度，若有相關再進一步判定抄襲程度。內、外部相關判定模組架構如圖 1、圖 4 所示:

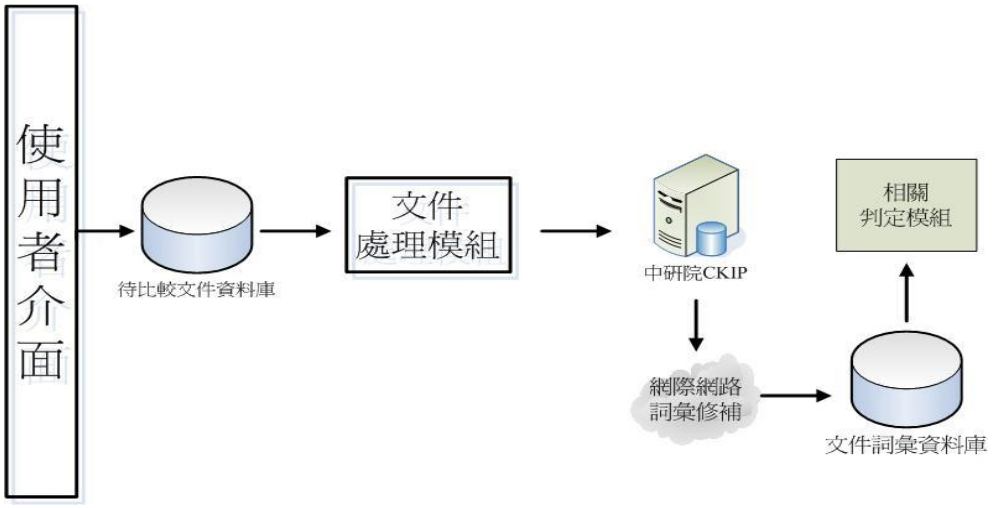

圖 1　內部相關判定模組架構

圖 1 為內部相關判定模組架構，所謂內部相關判定是讓使用者個別上傳兩個或兩個以上不同檔案分別去做比較，其中包括被比較文件資料庫、待比較文件資料庫、相關判定模組、抄襲判定模組等元件，內部相關判定元件說明如下:

（一） 待比較文件資料庫：此資料庫中的文件由使用者上傳，用來儲存與被比較文件的比較資料。

（二） 文件處理模組:用來處理省用者上傳後的文件，並傳給中研院 CKIP 進行斷詞。

（三） 中研院 CKIP:中研院 CKIP 是將一段句子透過網際網路傳入中研院 CKIP 進行斷詞，以便本系統處理句型架構。

（四） 網際網路詞彙修補:經由中研院 CKIP 斷詞後的詞彙透過 Google、維基所提供的 XML 查詢資訊進行詞彙修補，並將修補成功的詞彙存於資料庫當中，以減少人工建立的時間。

（五） 文件詞彙資料庫:本研究將每個使用者上傳的檔案進行剖析並儲存，以便快速配定是否相關。

（六） 相關判定模組：此模組結合了語意分析來判定文章相關程度，若有相關才會進行抄襲的判定，此模組詳細架構如圖三所示。

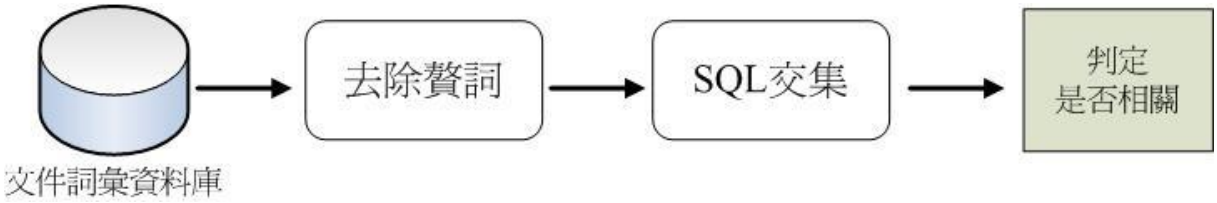

圖 2 相關判定模組架構

（一） 文件詞彙資料庫:本研究將每個使用者上傳的檔案進行剖析並儲存，以便快速配定是否相關。

（二） 去除贅詞:將不必要的詞或是常出現的用語進行剔除，例如:本研究、的、很、因此等常用語進行剔除，以降低判斷的失誤率。

（三） SQL 交集:使用 SQL 與法中的交集(INTERSECT)進行判定,若有相關才會進行抄襲的比對。

圖三為相關判定模組演算法。

```
public void Beginning(int FileParagraph)
{
  Find compare file number From Database and compute this file 'words' count

  Find be compared files  number From Database and compute this file 'words' count

  if Both of the above divided > 20 %
      then comparative Analysis

}
```

圖 3 相關判定演算法

圖 4 為外部相關判定模組架構，而所謂外部相關判定是讓使用者上傳一個或多個不同檔案分別與網際網路相關文件做比較，其中包相關文件資料庫、待比較文件資料庫、中研院 CKIP、文件處理模組、相關網頁處理模組、搜尋相關文件模組、等元件，相關元件說明如下:

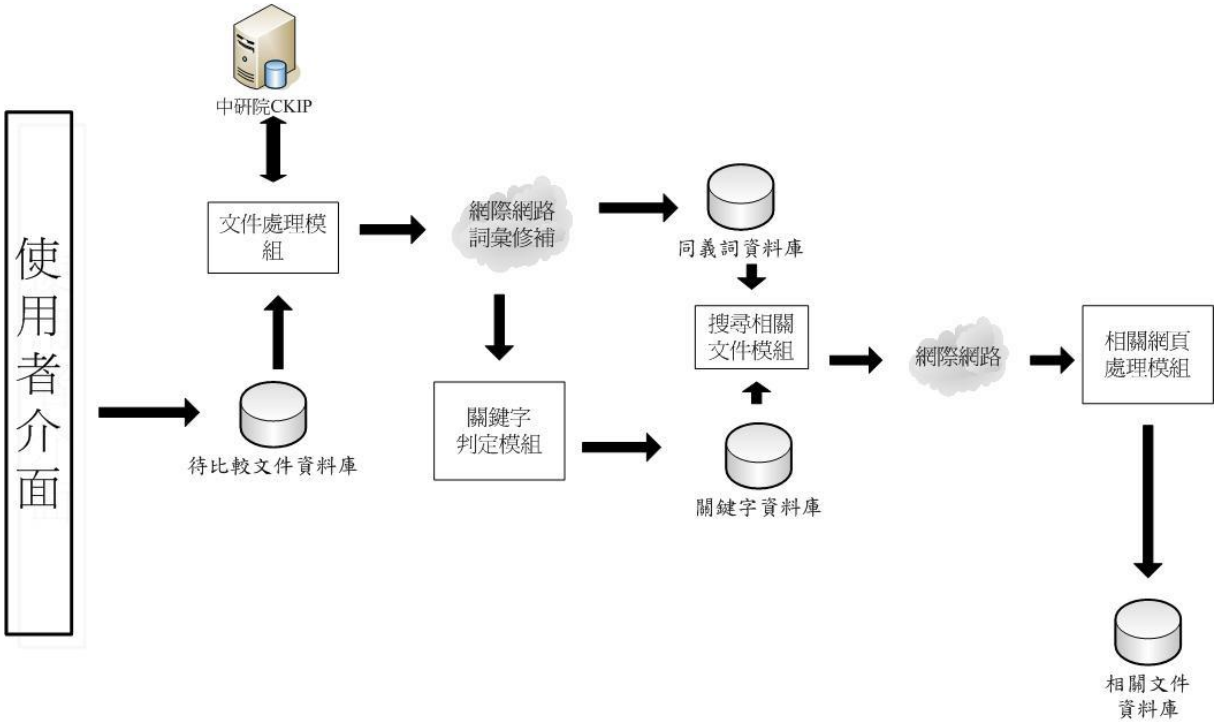

圖 4 外部相關判定模組架構

（一） 相關文件資料庫:此資料庫中的文件是透過網際網路搜尋相關文件，進行相關的判定，若有搜尋到相關文件則進行抄襲的比較。

（二） 待比較文件資料庫: 此資料庫中的文件由使用者上傳，是用來儲存與網際網相關文件比較的資料。

（三） 關鍵字資料庫: 用來儲存每個文件每段的關鍵字。

（四） 中研院 CKIP:中研院 CKIP 是將一段句子透過網際網路傳入中研院 CKIP 進行斷詞，以便本系統處理句型架構及關鍵字等。

（五） 文件處理模組:此模組處理使用者所上傳的文件，並傳給 CKIP 進行斷詞。

（六） 關鍵字判定模組:本研究使用每個詞在段落中所出現的詞頻來判定每段的關鍵字。

（七） 網際網路詞彙修補:經由中研院 CKIP 斷詞後的詞彙透過 Google、維基所提供的 XML 查詢資訊進行詞彙修補，並將修補成功的詞彙存於資料庫當中。

（八） 相關網頁處理模組: 由於各個資訊來源網站透過此模組，只保留需要之網頁資訊內容，然後將該資訊送至 CKIP 進行斷詞處理，以便比較判斷之用。

（九） 搜尋相關文件模組:此模組接收經過處理後的文件，進行網際網路上搜尋。

當本系統完成文件比對的相關分析後，若兩份文件或是檔案具有相關性，則將會由系統進行抄襲的明確判定，並提供抄襲的彙整資訊供教師或使用者參考。本研究提出的系統架構，如圖 5。

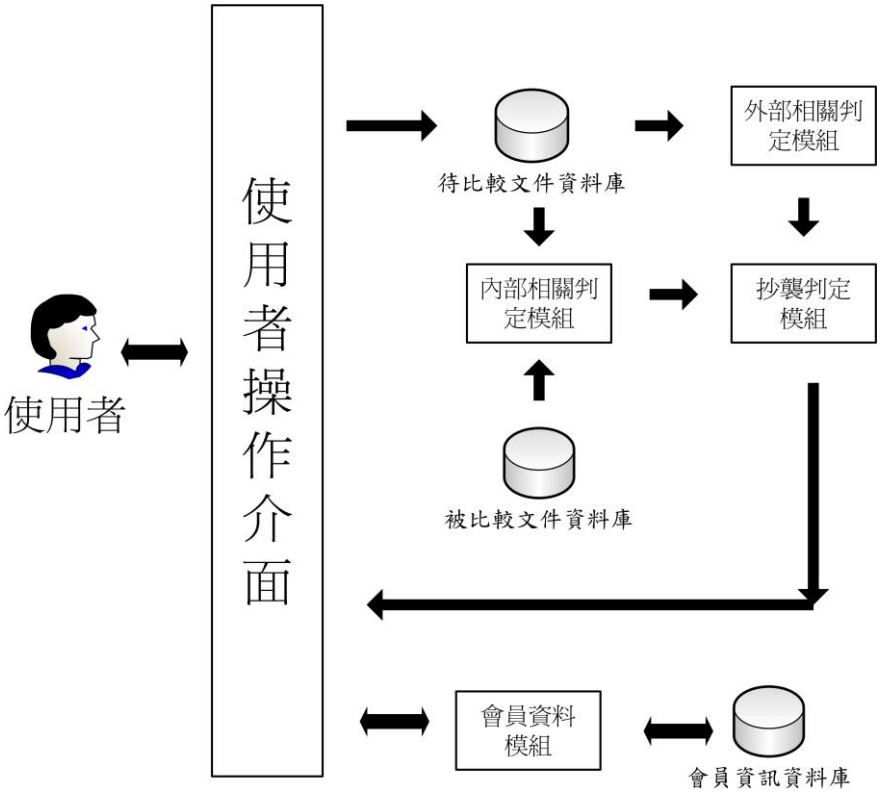

圖 5 系統架構

（一） 使用者操作介面:提供使用者操作系統的功能。

（二） 會員資訊資料庫:用來存取本系統會員基本資料。

（三） 被比較文件資料庫:此資料庫中的文件由使用者上傳，是用來儲存將做為被比較的資料。

（四） 待比較文件資料庫: 此資料庫中的文件由使用者上傳，是用來儲存與被比較文件比較的資料。

（五） 會員資料模組:若要使用本系統須註冊成為會員，以便管理此系統使用人員。

（六） 內部相關判定模組:此模組使用語意分析來判定文章相關程度，若有相關才會進行抄襲的判定。

（七） 外部相關判定模組:經由此模組可將使用者上傳文件進行網路上的搜尋，並將相關的網路資訊下載並進行比對。

（八） 抄襲判定模組:此模組用來比較文件與文件或文件與網際網路上的資訊是否有抄襲的可能。

四、 系統成效評估

4.1 雛型系統功能說明

以下僅針對雛形系統介面進行說明。

圖 6 為使用者上傳文件查詢，使用者可透過此介面進行上傳文件的查詢和刪除。

圖 6 文件查詢

使用者可透過文件上傳介面，進行兩種不同文件的比較，如圖 7 所示。

圖 7 文件上傳

使用者上傳文件後可透過比對結果介面進行查詢如圖 8 所示。

圖 8 比對結查詢

在文件比對前，會透過關鍵字進行相關文件的分類，比對結果會將相似部分以紅色字體標示如圖 9、圖 10 所示

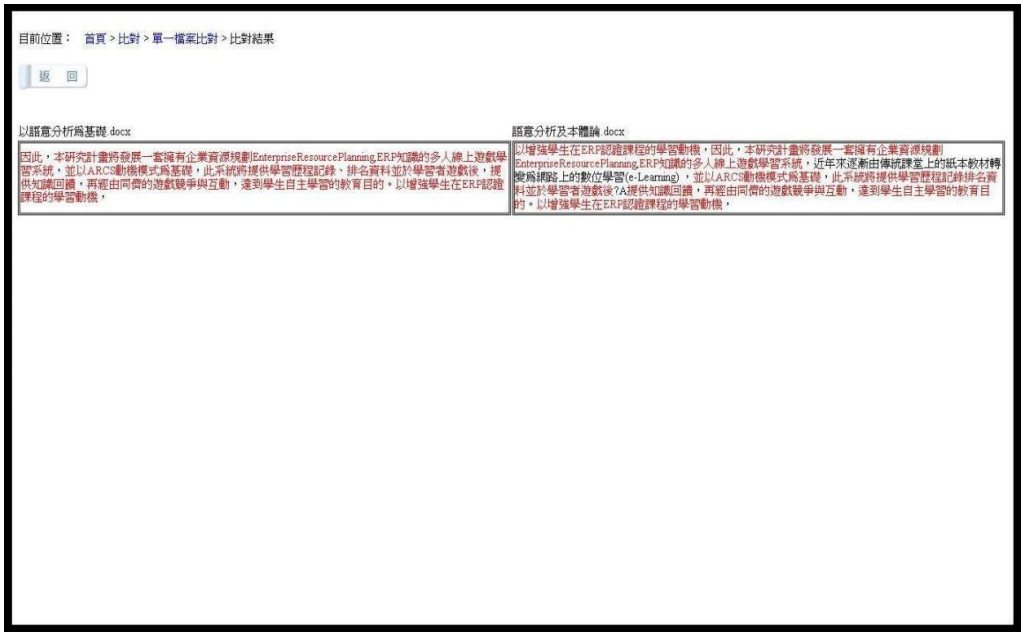

圖 9 內部比對結果

圖 10 內部比對結果

使用者亦可上傳單一文件進行外部比對，而系統會自動將比對到的相似的網站網址呈現給使用者如圖 11 所示。

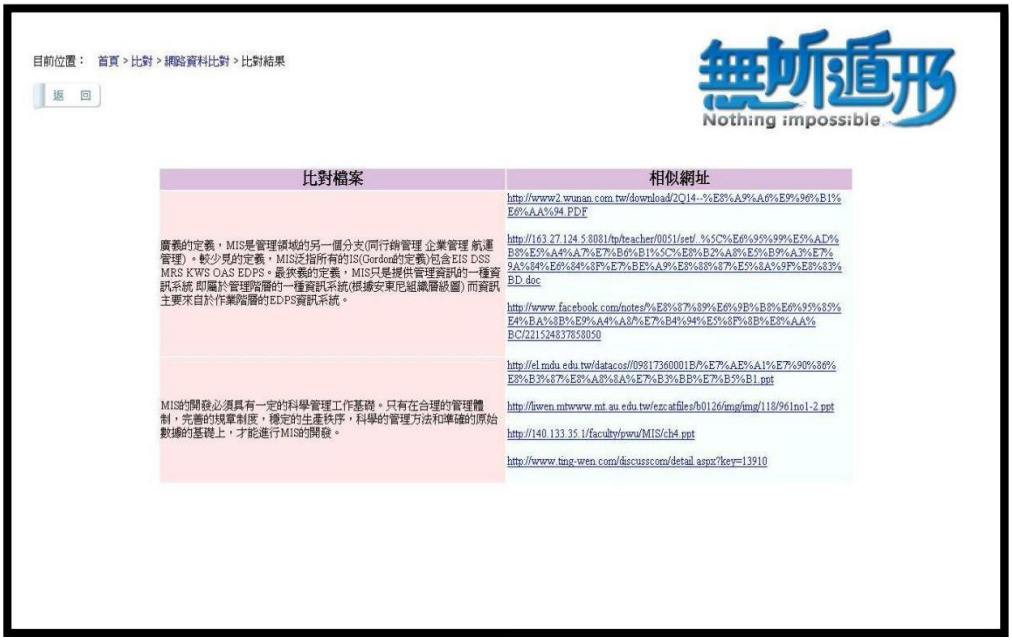

圖 11 外部比對結果

五、 結論與建議

本研究希望能藉由資料檢所及語意分析等技術，並且以結合網際網路的方式，建構出一個可以提供使用者參考的比對系統，期望藉由此系統杜絕抄襲的風氣。

5.1 研究貢獻

本研究藉由資訊檢索及語意分析技術讓抄襲無所遁形，本研究之貢獻，透過語意分析辭彙修補，事前將文件進行相關的分類增加比對的效率，以及透過網際網路的幫助省去過去需人工新增辭彙修補詞的時間，以利若有新詞出現亦可進行辭彙修補。

5.2 研究限制與未來研究建議

本研究在分析文件部分因資料量不足而造成的分析文件效率的降低，因此本研究後續將進行大量的資料建立以及改進現有演算法來提升文件分析的效率。本研究目前僅針對 WORD 檔案格式進行分析而未將其他檔案類型納入考量，因此後續研究會將其他檔案類型納入研究中。

5.3 計畫成果自評

本研究對於辭彙修補運用網際網路的便利性，進行線上辭彙修補省去大部分的人工建立時間，在比對文件方面當文字中插入不相關或是多於贅詞時亦能準確的比較出來，對於網路比對部分亦能將類似或相關網站呈現出來給予使用者參考。

六、 參考文獻

中央研究院資訊科學研究所(2004)，中文詞知識庫網站，http://ckip.iis.sinica.edu.tw/CKIP/。

李煜基、洪一梅(2007)，“相關與模糊在資訊檢索領域中關係驗證與分析”， 圖書與資訊學刊 ，60 期。

黃居仁(2003)，“語意網、詞網與知識本體，淺談未來網路上的知識運籌”，佛教圖書館館訊，第 33 期，頁 1-16。

應鳴雄、鄧光宏(2011)，”以本體論及語意分析為基礎的二階段藝文資訊整合技術”，第十七屆資訊管理暨實務研討會，地點: 台南，頁 1-12。

陳稼興、謝佳倫、許芳城(2000)，“以遺傳演算法為基礎的中文斷詞研究”，資訊管理研究，第 2 卷，第 2 期，頁 27-44.

鄭景俗、陳智賢、蘇勇戩(2008)，”基於模糊權重資訊檢索整合技術之推薦系統”，電子商務學報，10 卷 3 期。

Lancaster, F. W. and Warner, A. J. (1993), “Information Retrieval Today”, Arlington, Virginia:Information Resources press.

Mohammadian,M.(1999), “Computational intelligence formodeling, automation:Evolutionary computation & fuzzy logic for intelligence control,knowledge acquisition & information retrieval”,Netherlands:IOS Press,pp.176-194

Wixom, B. H., & Todd, P. A.(2005) “A Theoretical Integration of User Satisfaction and Technology Acceptance” Information Systems Research, (16:1) pp.85-102.

Zadeh, L.A.(1965), “ Fuzzy Sets,” Information and Control, Vol. 8, pp.338-353.

雲書苑教育科技(2012), “PPvS_org 在線論文抄襲比對系統”, Available at: http://www.ppvs.org/

Hong Kong School Net (2012), “VeriGuide introduction,” Available at: http://veriguide1.cse.cuhk.edu.hk/portal/plagiarism_detection/index.jsp.

PlagiarismDetect.com(2012), “Detection Premium Plagiarism Detection System,” Available at: http://www.plagiarismdetect.com/