方法式專利步驟相似度之比對

陳振原[1]、蘇豐文[1,2]

國立清華大學資訊系統與應用研究所[1]

國立高雄大學資訊工程系[2]

soo@cs.nthu.edu.tw;soo@nuk.edu.tw

摘要

此研究的主要目的為建立兩方法專利的自動相似度的分析的技術， 以減少此類型專利分析時的人物力消耗。 本系統提出由文法樹來進行特徵配對， 建立元件的相似度比較方法以及介係詞的相似度比較方法 。相似度比對方法並儘量符合心理學與法律上判斷原則。我們從專利申請範圍中自動擷取出各項元件、元件之間的關係、步驟、步驟之間的關係，而結構化所擷取物件建成樹狀機讀式結構，進行相似度比對以避免侵權。本研究進行兩實驗以評沽效益: 1.利用基因演算法 以學習步驟相似度計算中的權重參數值，2. 與一般以統計文字出現頻率的相似度算法 BLEU 進行比較。實驗結果本系統提供的方法較自然語言常採用的BLEU 方法更適合應用在專利步驟的相似比對上。

關鍵詞: 專利侵權比對，專利範圍，方法專利，機讀式檔案，步驟相似度比對

一、緒論

目前專利分析大多是以宏觀的角度切入得到大量相關專利的統計分析圖表， 較少對兩篇專利文件的內容作詳細比較分析。 但是兩篇專利文件之間的比較分析往是侵權判斷與迴避的重要關鍵。以往專利工程師以人工方式進行分析，耗費大量人力物力，為企業的一大成本。 專利範圍解讀錯誤! 找不到參照來源。是判斷侵權流程錯誤! 找不到參照來源。的第一步驟，此過程必須經過專利工程師解讀完專利申請範圍後自行寫出主要元件、次要元件、所有元件間的關係、步驟，以利爾後的全要件原則比對、均等論或是消極均等論的比對。 因為可專利標的可歸納分為物、方法、用途三種，所以針對不同發明標的的專利文件有不同的解讀方向。 物標的與方法標的的專利文件解讀方向不同，方法標的專利主要描述產品的製造方法或是無產品的技術方法，而物標的的專利主要描述物的結構。國立清華大學人工智慧實驗室已研發出一套系統錯誤! 找不到參照來源。將物標的之專利結構化，從而進行專利範圍解讀以及協助侵權判斷流程中全要件原則比對。物標的的專利文件可經由此系統轉換成機械式可讀的樹狀結構。本研究欲建立一套系統，使其能自動並進行兩方法標的專利中的步驟和方法相似度的分析，以減少此類型專利分析時的人物力消耗。如專利標號 6979646 方法標的專利之第一條申請範圍(圖 1) 可轉換成結構化的機械式可讀結構(圖 2)，結構中包含：

根節點：如圖 2 的“A method”； **功能節點：**經由“for”關係與根節點相連，且敘述根節點之功能的節點。如圖 2 的“forming hardened interconnects”;**步驟節點：**經由包含關係(comprise、 have、 include etc.)與其他節點相連的節點。例如圖 2 的“depositing a metal film over a semiconductor wafer surface”; **步驟間的關係：**即步驟節點之間的關係，敘述兩步驟之間的實施順序(after、before、during、

while etc.)或是包含關係。例如圖 2 的

6,979,646

1. A method for forming hardened interconnects comprising: depositing a metal film over a semiconductor wafer surface; introducing an additional metal species comprising beryllium to the metal film; heating the deposited metal film with the introduced metal species; allowing the metal film to cool, so as to form precipitates of the introduced metal species; and after allowing said heated metal film to cool performing chemical-mechanical polishing wherein the additional metal precipitates harden said deposited metal film to reduce the rate of said polishing.

圖 1　6979646 號方法標的專利之第一條申請範圍

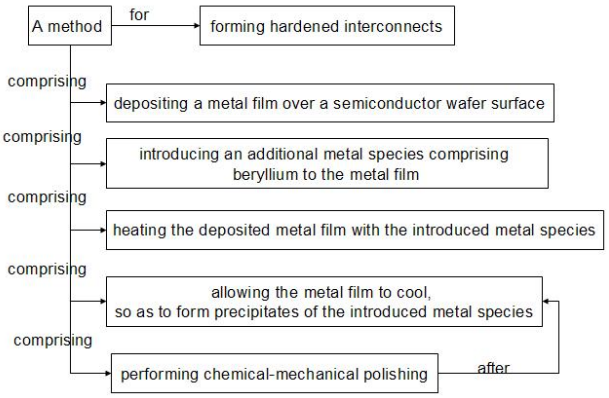

圖 2　6979646 號專利之第一條申請範圍的方法流程圖

二、研究架構與方法

2.1 前處理：

1.　從取得的專利文件中切割出專利號碼、標題、摘要、申請範圍(claims)、說明書(description)匯入資料庫。 2.專利文件的申請範圍敘述中第二次提起的事物使用 said 替代 the 定冠詞。3. 將專利文件中的申請範圍依照條目編號將每一條申請切割，並且將所有空白字元取代成一格空白。4.還原不應該分開的符號，例如縮寫符號(vs.=> vs.)、 破折號(polishing - pad => polishing-pad)、小數點(5. 23 => 5.23)等。

2.2 擷取元件

1.利用 Stanford tagger **錯誤! 找不到參照來源。**將資料庫的中每條申請範圍中詞彙皆標上適當詞性。

2. **多字詞(N-gram)分割,** 根據統計，一般元件的字數不會超過五個字，所以將各條申請範圍分割成 1-gram、2-gram、3-gram、4-gram、5-gram。

3.**刪除含有停用字(stop word)的多字詞**，可減少判斷符合元件詞性的數量,進而加快處理速度。

4.**刪除不符合元件詞性的多字詞**, 過濾掉具有 stop word 後的多字詞，利用正規表示式“(VBG|VBN|JJ|JJR|JJS|RB|RBR|RBS|NN|NNS|NNP|NNPS)*(NN|NNS|NNP|NNPS)”來比對挑選符合此正規表示式的多字詞。正規表示式中各符號*為 Stanford tagger 中的詞性*。

5. **篩選元件**

根據書寫習慣，元件第二次被提起時會省略多餘的形容詞修飾，所以經過步驟 3 處理過的多字詞若在申請範圍裡的前一個字是 the，則將其視為一個元件。 而剩餘的多字詞將經過統計出現頻率的方法刪除不到門檻值的多字詞。

2.3 功能與步驟的擷取

功能的定義為方法專利提起的目的，也就是製造一個產品或是達到某種效果，例如“forming hardened interconnects”即是 6979646 號方法標的專利的功能敘述。步驟的定義為欲達成專利所描述的產品製造方法或是無產品的技術方法而實施之操作或處理動作。例如“depositing a metal film over a semiconductor wafer surface”即是一個步驟。

本研究利用 Stanford parser **錯誤! 找不到參照來源。**來對申請範圍作文法剖析，再從剖析後的結構中抽取出功能與步驟。唯專利文件本身文法結構獨特，用字遣詞與文法結構與一般自然語言文件不同，且存在著隨不同領域而有不同冷僻的用字。 可利用取代元件為”entity#entity”讓 parser 較方便剖析，避免多字詞中的某個字眼詞性被誤判，而造成文法剖析的錯誤。文法結構獨特問題本研究提出下列

前處理方法：因為獨立項的申請範圍包括前言、連接語及主體三部份，所以利用連接語將獨立項申請範圍切割成前言以及主體兩半，並將連接詞取代成適當的動詞型態,例如“comprising” 取代成“comprises”，再丟入剖析器分析文法。從剖析後的結果裡擷取出該專利所擁有的功能。並將前言刪除功能敘述以縮短句子長度，增加剖析的速度與減少剖析的錯誤率。例如 6979646 號專利(圖 1)的前言“A method for forming hardened interconnects comprising ”經過替換連接詞後的剖析結果為圖 3， 圖中“entity1entity”為元件“hardened interconnects”被替換為“entity#entity”之後剖析的結果。

```
det(method-2, A-1)
nsubj(comprises-6, method-2)
prep(method-2, for-3)
dep(for-3, forming-4)
dobj(forming-4, entity1entity-5)
```

圖 3 6979646 號專利之第一條申請範圍的前言剖析

利用圖 3 中 Stanford parser 提供的從屬性**錯誤! 找不到參照來源。**（圖 5）“prep”、“dep”、“dobj”，可截取出功能敘述 “for forming entity1entity”。

利用標點符號切割主體成數個短句，並一一循序接在先前取代後的前言之後，再丟入剖析器剖析，例如 6979646 號專利(圖 1)經過替換元件等處理後變為“A method for forming entity1entity comprises depositing a entity2entity over a entity3entity …”，剖析器剖析的結果為圖 4，利用 Stanford parser 提供的從屬性(如圖 5) “partmod”、“dep”、“dobj” 、“prep”、“pobj”即可截取出步驟敘述“depositing a entity2entity over a entity3entity”，再還原替換的元件名稱則可得“depositing a metal film over a semiconductor wafer surface”。

```
det(method-2, A-1)
nsubj(comprises-6, method-2)
prep(method-2, for-3)
dep(for-3, forming-4)
dobj(forming-4, entity1entity-5)
partmod(comprises-6, depositing-7)
det(entity2entity-9, a-8)
dobj(depositing-7, entity2entity-9)
prep(depositing-7, over-10)
det(entity3entity-12, a-11)
pobj(over-10, entity3entity-12)
```

圖 4 6979646 號專利前言後一步驟的剖析結果

dep - dependent	ref - referent
aux - auxiliary	expl - expletive (expletive there)
auxpass - passive auxiliary	mod - modifier
cop - copula	advcl - adverbial clause modifier
conj - conjunct	purpcl - purpose clause modifier
cc - coordination	tmod - temporal modifier
arg - argument	rcmod - relative clause modifier
subj - subject	amod - adjectival modifier
nsubj - nominal subject	infmod - infinitival modifier
nsubjpass - passive nominal subject	partmod - participial modifier
csubj - clausal subject	num - numeric modifier
comp - complement	number - element of compound number
obj - object	appos - appositional modifier
dobj - direct object	nn - noun compound modifier
iobj - indirect object	abbrev - abbreviatio
pobj - object of preposition	advmod - adverbial modifier
attr - attributive	neg - negation modifier
ccomp - clausal complement with internal subject	poss - possession modifier
xcomp - clausal complement with external subject	possessive - possessive modifier ('s)
compl - complementizer	prt - phrasal verb particle
mark - marker (word introducing an advcl)	det - determiner
rel - relative (word introducing a rcmod)	prep - prepositional modifier
acomp - adjectival complement	sdep - semantic dependent
agent - agent	xsubj - controlling subject

圖 5 Stanford parser 的從屬性關係

2.4 建立方法流程圖

所謂方法即為一篇方法專利中描述的產品製造方法或是無產品的技術方法，也就是所有步驟的集合，且步驟間具有被實施的順序關係。基於方法的定義而制定的方法流程圖架構需先擷取出功能、步驟及步驟間的關係，再依照其相對應的關係聯結成方法流程圖。利用類似取代元件的方法將擷取出的步驟替換成“step#step”並建立參照關係。例如 6979646 號專利的第一條申請範圍(圖 1)中敘設的“allowing the metal film to cool, so as to form precipitates of the introduced metal species”與“and after allowing said heated metal film to cool performing chemical-mechanical polishing ”中的“allowing said heated metal film to cool”顯然是同一個步驟，但是在擷取步驟的階段並不會發現是相同的步驟，導致替換成“step#step”時會替換為不同的流水號。故要建

立參照關係，將該二步驟視為相同的步驟。建完參照關係後讓剖析器擷取出步驟與步驟間的關係生方法流程圖。

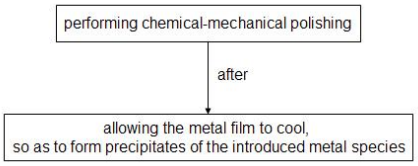

圖 6 三元體示意圖

2.5 步驟的特徵組成及配對方法

專利申請範圍中的步驟的主體為動名詞加上受詞，故步驟寫法在符合英文文法下，書寫類型的變化有限，通常僅僅是加上從屬子句作補充說明，如關係子句，形容詞子句，副詞子句。這些子句是文法上的結構定義，在文法樹上會自然而然的剖析到適當的子樹。例如由“since”帶頭的副詞子句在文法剖析樹中會分配到(SBAR(IN since))標籤下，而由 who 開頭的形容詞子句在文法剖析樹中會分配到(SBAR WHNP(WP who))。本系統提出由文法樹來進行特徵配對，讓比較的特徵屬於同一特徵分類。步驟的組成特徵定義為經由剖析器剖析出的文法結構樹中的每個葉節點，特徵類別同時考量詞性及兩剖析樹中相對應的位置，如圖 7 為“measuring the weight of a slurry tank for preparing the slurry with a weight sensor located on the slurry tan”和“*measuring an additive concentration in the slurry in a tank for preparing the slurry with a concentration sensor located in a closed loop pipe*”兩步驟的剖析結果，唯步驟事先加上了主詞(I)並將開頭動名詞還原成原型，以及元件替換成流水號以避免剖析器剖析錯誤。

```
(ROOT
  (SQ
    (NP (PRP I))
    (VP (VB measure)
      (NP
        (NP (DT the) (NN E_8))
        (PP (IN of)
          (NP (DT a) (NN E_2))))
      (PP (IN for)
        (S
          (VP (VBG preparing)
            (NP (DT the) (NN E_1))
            (PP (IN with)
              (NP
                (NP (DT a) (NN E_3))
                (VP (VBN locatcd)
                  (PP (IN on)
                    (NP (DT the) (NN E_2)))))))))))
```

```
(ROOT
  (SQ
    (NP (PRP I))
    (VP (VB measure)
      (NP
        (NP (DT an) (JJ additive) (NN E_7))
        (PP (IN in)
          (NP
            (NP (DT the) (NN E_1))
            (PP (IN in)
              (NP (DT a) (NN E_5))))))
      (PP (IN for)
        (S
          (VP (VBG preparing)
            (NP (DT the) (NN E_1))
            (PP (IN with)
              (NP
                (NP (DT a) (NN E_6))
                (VP (VBN located)
                  (PP (IN in)
                    (NP (DT a) (NN E_4))))))))))))
```

的三元體(triple)(圖 6)，合併組這些三元體後再加上跟節點與功能的三元體，即可產

圖 7 兩步驟的剖析樹狀圖

圖 7 可經由樹中的第一組 VP,NP 節點去 travel 中步驟的主要作用動詞以及主要受詞，其餘葉節點則視為一般特徵，惟 DT 節點皆屬於停用字，本身對步驟內容並無影響，故不將其視為特徵。特徵配對的演算法中如屬 NN 系列(NN,NNS,NNP,NNPS)者皆視為相同，而圖 7 中的兩個剖析樹在此演算法下可產生圖 8 中的 12 個配對，其中主詞 I 為自行添加的，故可忽略之。

```
measure : measure
E_8 : E_7
of : in
E_2 : E_1
for : for
preparing : preparing
E_1 : E_1
with : with
E_3 : E_6
locatcd : located
on : in
E_2 : E_4
```

圖 8 圖 7 中兩步驟的配對圖

2.6 特徵的相似度計算

本系統主要採用 WordNet [11]的階層關係來做相似度計算， WordNet 的資料庫設計雖已收藏十五萬不同的字和建立二十餘萬組字與字的配對資料，但依然會缺乏專業術語的相關字。專業術語大抵分為動詞、名詞、形容詞、副詞，又最常出現的冷僻用法多半為名詞的專業術語。本論文研究收集的 1400 篇化學機械研磨專利中，動詞在 WordNet 查詢不到的比例為 3.2%而副詞在 WordNet 查詢不到的比例為 1.3%。因兩者的失誤率 (miss rate) 並不太高。專業術語中的名詞和形容詞絕大多數在本系統的前處理中已經替換成元件，且元件多半是二到五字的組合字，導致失誤率 (miss rate)高達 88.05%，例如化學研磨領域中很常出現的元件“polishing pad”就查詢不到。 因 WordNet 沒有介係詞的資料，故本系統必須建立元件的相似度比較方法以及介係詞的相似度比較方法。針對單一特徵的比對模式可分為三大類：一

般字類，介係詞類，元件類。

一般字類即非介係詞和元件的其他字眼，相似度由 WordNet 所提供的階層關係來計算，WordNet 提供的查詢中與相似度有關的屬性有同義、相似、上層(Hypernym)、下層(Hyponym)，分別定義相同字及同義字的相似度為 1、相似字的相似度為 0.9、上層字的相似度為 0.6、下層字的相似度為 0.6，其中相似度 1 表示完全相同，相似度 0 表示完全不同。此處定義數值為 1、0.9、0.6 的目的在於區別出特徵的相似度高低，數值僅侷限於同義字定義值須大於相似字定義值，且相似字定義值須大於上下層字定義值。步驟相似度是由各個特徵的相似度分數加乘而來，所以改變一般字類的相似度定義值，會相對改變步驟整體的相似度值。但此舉並不影響判斷兩步驟是否相似，因為當步驟整體相似度下降時，實驗中判斷相似與否的門檻值也會跟著下降。

計算兩個一般字類的相似值時，先任取一字和該字的詞性於 WordNet 查詢與另一字是否有上述的關係資料，若存在的話，依上述關係定義給定相似值。例如要計算動詞“measure”和“determine”的相似度，WordNet 中“measure”的 Synset id 有十二個，但詞性為動詞的“measure”的釋義僅有四種，分別編號為 200476505、200626409、 200658664、202623026，由這四個編號在 WordNet 中依序從同義、相似、上層、下層關係中找尋是否和“determine”有相關聯，最後由編號 200626409 於上層關係中找到編號 200674417 的“determine”，故得知 “determine”是“measure”的上層字。依照前述定義，“measure”和 “determine”的相似度為 0.6。相似度依階層相差數等比遞減，例如編號 200116175 的“let”為編號 200115695“make”的下層字，且該“make”為編號 200121430 “change”的下層字，由於每層上下層字的相似度定義為 0.6，故 “let” 與 “change” 的相似度值為 0.6*0.6=0.36。

本系統將專利文件中的介係詞分為十一類(表 1)，同類別的介係詞相似度定義為 1，不同類別的介係詞和未收錄於下表的介係詞相似度則定義為 0。例如“down”和“below”位於“在…下”的類別，所以“down”和“below”的相似度為 1。反之，“down”和“up”因為於不同類別，所以相似度為 0。

表格 1 介係詞分類表

	介係詞
在…上	"up","upon","on","onto","over","above","atop"
在…下	"down","below","under","beneath","underneath"
在…內	"in","inside","within"
在…外	"out","outside"
在…附近	"near","about","around","beside"
在…之中	"among","amid","between"
穿過	"through","throughout","across"
在…期間	"in","within","during"
在…以前	"before"
在…以後	"after","behind"
直到	"until","pending"

兩元件相似度值分為內部特徵相似度與外部特徵相似度。內部特徵有屬性(attribute)、本體(essence)兩類，而外部特徵有從屬(subordinate)、組成(composition)、位置(position)三類。

元件的本體特徵表示該元件的本質是何物，而屬性特徵則是修飾本體的辭彙，表示該元件的功能或狀態，元件的內部特徵相似度算法為屬性相似度和本體相似度的平均，本體和屬性則視為一般字類公式如下:

假設兩元件為 A 和 B：Ai 表示 A 元件的第 i 個屬性，#A 表示 A 的屬性個數， #A ≤ #B

$S_{essence}$=wordnet(essence of A, essence of B);

$$S_{attribute}=\begin{cases}0, if\ \#A=0\\ (\sum_i \max(\text{wordent}(A_i, \text{all Bj})))/\#A\end{cases}$$

$S_{inner\text{-}feature} = (S_{attribute} + S_{essence})/2$

專利範圍中元件必定透過某種關係與另一元件相連，該關係與任一元件則為另一元件的外部特徵，也就是專利範圍解讀後所拆解出的三元體（元件 A，關係 R,元件 B)中，元件 A 的外部特徵為(關係 R,元件 B)，元件 B 的外部特徵為(元件 A,關係 R)。

外部特徵種類列於下表(表 2)，主要分為 A.從屬、B.組成、C.位置三類。從屬類可細分為 1.至少包含，2.組成元素，3. 主要組成元素，4. 包含 [10]四類。至少包含法律界定為開放式的權利保護範圍，權力範圍包含以下所述元件，但不排除包含其他元件。組成元素法律界定為閉鎖式的權力保護範圍，權力範圍僅包含所列舉的元件，增加或減少一個元件便可迴避權力涵蓋範圍。主要組成元素法律界定為介於開放與閉鎖式之間的權力保護範圍，其權力範圍除了所述的元件，但不排除一些次要的元件。此寫法的權力範圍涵蓋亦介於兩者之間。包含為除了上述法律給予特定解釋的三種從屬類外其餘的從屬關係，須經由法院的裁定才能確定其保護的權力範圍。組成類在專利中表示組成或是形成的慣用法，通常使用“formed *IN*”跟“composed *IN*”型態表示，其中 *IN*代表可能接在“formed”或“compose”之後的介係詞。位置類在專利中描述元件間的位置關係。大抵分為相鄰(Adjacent)，環繞(Surround)，相對(Opposite)，黏附(Adhered *IN*)，嵌入(Embedded *IN*)，位於(Positioned *IN*)六類，其中 *IN*代表可能出現的介係詞。

表格 2 兩元件間的關係種類

關係類別		關係名稱
包含	至少包含	"comprise(s)","comprising"
	組成元素	"consist(s) of", "consisting of"
	主要組成元素	"consisting essentially of"
	包含	"include(s) ","including"
		"have", "has", "having"
		"contain(s)","containing
		"
組成	Formed *IN*	以下介係詞前皆省略“formed”。 "above", "along", "around", "as", "at", "atop", "between", "by", "during", "for", "from", "in", "inside", "into", "of", "on", "over", "through", "to", "under", underneath", "upon", "with", "within", "without"
	Composed *IN*	以下介係詞前皆省略“composed”。 "at", "from", "of"
位置	Adjacent	"adjacent", "adjacent to"
	Opposite	"opposite", "opposite to"
	Surround	"surrounding"
	Adhered *IN*	"adhered on", "adhered to", "bonded to""attachable to" , "attached to"
	Embedded *IN*	以下介係詞前皆省略“embedded”。 "at", "in", "into", "within"
	Positioned *IN*	以下介係詞前皆省略“positioned”。 "above", "across", "against","along", "around", "as", "at", "below", "beneath", "beside","between", "beyond", "by", "during", "for", "from", "in", "inside", "near", "of", "on","onto", "outside",

		"over","to", "towards", "under", "with", "within"

由於元件和元件之間都以某種關係互相關聯著，所以類似的元件就極可能以某種類似的關係連結著其他類似的元件。基於物以類聚的觀念，外部特徵的相似度算法以統計大量的相同領域專利文件中的元件三元體，藉以歸納出元件之間的相似度。本系統從 1400 篇化學機械研磨專利中擷取出含有表 2 所列舉的關係的三元體共 43286 條(20604 種)，再經過同義和反義關係的修正後剩 19811 種不同的三元體。

經由上述處理過的統計資料，本系統將元件的所有外部特徵相似度表示為多維向量，該向量代表元件具有向量中的各項特徵。在心理學的特徵模式中，兩物的相似程度是採兩者所有的全部特徵來進行比對，故本系統聯集兩元件的所有外部特徵值，以利之後兩元件相似度的運算。在統計資料中可求得每種外部特徵的出現次數，而出現次數越高的外部特徵表示元件與此外部特徵兩者極為相關，所以外部特徵的重要性應該隨著出現次數越多而越重要。本系統利用外部特徵次數除以該元件出現總次數來作正規化，如此可區別出每個外部特徵的重要性，而正規化後的向量也表示出每項特徵的權重。

權重化外部特徵座標是個數量向量，數量向量間的相似度計算大抵分為明可斯基距離(Minkowski Distance)和角度分離量(Angular separation)兩種。明可斯基距離是一般化的距離公式，所求得的距離必定大於零，當距離越近時兩點的相似度則越高。假設在m維空間中有兩點X=(X1, X2, …, Xm)，Y=(Y1, Y2, …, Ym) ，則其明可斯基距離可表示為：

$$D_{X,Y} = \sqrt[P]{\sum_{K=1}^{m} \left| X_K - Y_K \right|^P}$$

當P=2 時即是常見的歐幾里得距離(Euclidean Distance)，P=1 跟P=∞則為城市街道距離 和柴比雪夫距離(Chebyshev Distance)。

角度分離量也就是餘弦相似度其以兩個多維向量間的角度差異來度量向量間的相異性，所得數據介於-1 到 1 之間，當兩向量角度越相近時，所求出的餘弦距離越接近 1，也就是兩向量越相似；反之，則越接近-1，也就是兩向量越相異。假設在 m 維空間中有兩向量 X=(X1, X2, …, Xm)，Y=(Y1, Y2, …, Ym) ，則其餘弦相似度可表示為：

$$S_{X,Y} = \frac{X \bullet Y}{\|X\|\|Y\|} = \frac{\sum_{K=1}^{m} X_K * Y_K}{\sqrt{\sum_{K=1}^{m} X_K^2 * \sum_{K=1}^{m} Y_K^2}}$$

由於明可斯基距離求得的距離僅最小值有下限為零，但最大值並沒有上限，此種數值表現意義僅能表示出兩點完全相同，卻無法顯示出兩者完全無關的情形，且直到最大值出現之前，無法正規化到一般常用的相似度範圍 0 到 1 之間。而角度分離量在數學上的定義即為兩向量的餘弦值，餘弦值本身即位於-1 到 1 之間，加上本系統定義的權重化外部特徵向量中每一個數值皆大於等於零，所以在向量的限制下，整個餘弦值自然介於 0 到 1 之間 (0 表完全無關，1 表示完全向同)，故本系統採用角度分離量來計算權重化外部特徵向量。

2.7 步驟的相似度計算

方法專利的內容是敘述如何達成某種功能或是製造出某種物體的方法，方法中列出一系列步驟和實施流程，一但別人以相同方式達到相同目的則可依專利權進行控訴要求賠償。方法專利描述的重心是步驟和步驟的執行順序，所以判斷是否侵害到別人的專利時，步驟的相似程度是一個很重要的考慮點，故本系統提出一套步驟相似度的計算方法，冀望能儘量符合法律上判斷原則。

步驟的相似度基於心理學中的特徵模式設計，而步驟僅描述如何對某物執行動作，不考慮是由何者何物來執行此步驟，所以步驟並不會有主詞的特徵。相似度的算法則依各種特徵來各別進行相似度計算後，再給予權重化後相加得之。

步驟的特徵計算公式如下：

$S_{step} = a * S_{verb} + b * S_{object} + (1-a-b) * S_{other\ features}$

其中

- $0 \le S_x \le 1$, for all x

- $0 \le a, b, (1-a-b) \le 1$
- a, b, (1-a-b)代表 S_{verb} 、S_{object} 、$S_{other\ features}$特徵項相似值的權重。
- S_{verb}代表作用動作的相似度，算法為利用WordNet 的一般字類算法。
- S_{object}代表受作用者的相似度，算法為考慮內外部特徵的元件類算法。
- $S_{other\ features}$ 代表其餘特徵的相似度，所謂其餘特徵是步驟特徵配對後，非動詞與受詞的其餘特徵。相似度的算法採用Jaccard's Coefficient。此法概念是用兩集合的交集數量除以聯集數量，公式為

$$S_{AB} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$，由於步驟的其餘配對特徵可視為一個集合，故本系統在Jaccard's Coefficient概念下定義

$$S_{other\ features} = \begin{cases} 0 & ,if\ |A| = |B| = 0 \\ \dfrac{\sum S_{matched\_feature}}{|A| + |B| - \sum S_{matched\_feature}} & ,otherwise \end{cases}$$

其中 $S_{meached\_feature}$ 是每個配對特徵的相似度值，按特徵所屬類別(一般字類、介係詞類、元件類)設計的相似度計算方法求得。

三、實驗

實驗用資料的來源為 USPTO 的資料庫中[1]，對查詢範圍 “Title” 進行關鍵字 “chemical mechanical polishing” 或 “chemical mechanical planarization” 查詢到的 1400 篇化學機械研磨專利，以及由查詢範圍 “Title” 搜尋 “chemical mechanical polishing” 且 “method” 後找到的 428 篇化學機械研磨方法專利，本研究進行下列實驗：

- 基因演算法[11] 訓練步驟相似度計算的權重參數
- 與一般統計文字出現頻率的相似度算法進行比較

3.1 基因演算法訓練參數

本實驗從428篇化學機械研磨方法專利的申請範圍中選取50句步驟來進行步驟的權重參數訓練，參數除了步驟相似度公式本身定義的a, b, (1-a-b)之外，額外加入一個門檻參數，此參數表示專家認為兩步驟相似的最低相似值。而針對本實驗設計的各項基因演化機制如下：

- 基因編碼方式：隨機產生一百組染色體當作族群，而每個染色體具有四個基因，分別是a, b, (1-a-b), threshold。基因的編碼採用介於0到1之間實數如圖9

基因：	a	b	1-a-b	threshold
實數編碼：	0.5	0.3	0.2	0.3

圖 9 基因示意圖

- 挑選機制：採用混合取樣的精英挑選(Elitism)將適應函數前五十名的本代染色體直接複製至下一世代，以避免較適合的染色體在隨機取樣下(stochastic sampling)沒機會將基因留到下一代。而下一代的另五十個染色體則由本代的所有染色體隨機交配產生。
- 交配機制：因為染色體的基因編碼有介於0到1間的限制，且前三個基因的編碼值和必為1，所以不能採用傳統利用交配點來切割跟交換基因產生新子代的方法。為了符合基因的限制以及企圖朝適應的方向演化，定義交配出的子代各基因為雙親的平均值。
- 突變機制：設定交配後的染色體有5%的機會隨機挑選突變的基因，突變值為增加或減少基因值0.1。在基因編碼值的限制下，若突變後的基因值大於1則以1計，小於0則以0計。若突變的基因是前三基因之一的話，則在三者中隨機選出一基因來做對應的加減以維持前三基因編碼和為1。
- 適應函數：適應函數是計算步驟相似度求得的相似序列與專家提供的相似序列之間的各個同標地

的序列平均編輯距離。編輯距離是指一字串要轉換成另一字串所需要的最少編輯操作次數，因此適應函數值越小則兩群相似序列越相近，也表示該染色體越符合需求。

經由上述基因演算法的各機制設計方式進行一萬代演化，最後得到最佳的基因為圖 10，其適應值為 1.16，也就是說平均每個步驟的相似序列與標準相似序列相差一點多編輯距離。而其演化圖列於圖 11。

a	b	1-a-b	threshold
0.44105803	0.43911639	0.11982557	0.38122008

圖 10 基因演算法演化出的最佳基因

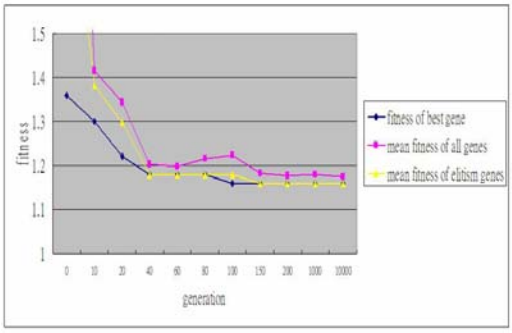

圖 11 適應值演進圖

平均編輯距離的衡量標準比準確率(precision)、召回率(recall)多了表現出數列排列順序的不同，但數值的表現僅知道當距離為 0 時表示完全吻合，以及數字越低表示越相似，並無法藉由一個距離數值來直接判斷準確與否。平均準確率 0.938、平均召回率 0.913。

3.2 與BLEU比較

BLEU[12]是一套評估機器翻譯品質的演算法，評估優劣的原則是依機器翻譯句和專家翻譯句的接近程度而定，越相似於專家翻譯句則該機器翻譯品質越優。因此 BLEU 有提供句子相似度的計算方法。本實驗將經基因演算法最佳化參數後的步驟相似度計算方法，與 BLEU 比較在方法專利中的步驟相似度的判斷準確度。

BLEU 演算法主要是計算候選句中的與標準答案相同的 N-gram 次數除以候選句中的 N-gram 的總數，當然 N-gram 出現次數越多，表示在 N-gram 下句子越相像。

$$P_n = \frac{\sum_{C \in Candidates} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in Candidates} \sum_{n-gram' \in C'} Count(n-gram')}$$

在 P_n 中即使句子過短導致分子變小，可是分母卻也同時變小了，所以並無法反應過短句與長句的差異性。BLEU 設有一簡潔處罰值(Brevity Penalty)以處罰機器翻譯句過短於專家翻譯句，機械翻譯句過長本身已在 P_n 中得到處罰，故並不特別設計處罰值。BLEU 的比較方法是單向拿機器翻譯句去和專家翻譯句比較，所以不需要有對稱性，也就是說 $BLEU_{A,B}$ 未必相等於 $BLEU_{B,A}$。但本系統處理的步驟相必須具有對稱性，故在不違反 BLEU 精神下，將 BP 值修改為

$$BP = \begin{cases} e^{(1-c/r)}, if\ c > r \\ e^{(1-r/c)}, if\ c \le r \end{cases}$$

最後整個 BLEU 相似計算公式如下，

$$BLEU = BP * \exp(\sum_{n=1}^{N} W_n * \log P_n)$$

其中 Wn=1/n，且 $n \in \{1,2,3,4\}$。

由於 BLEU 僅表達兩句句子的相似度，尚缺少判斷相似的門檻值，所以先採用基因演算法進行門檻值的訓練。僅調低突變值為 0.01 以及改用 BLEU 為兩步驟相似度的計算方法，實驗後得到的最佳門檻基因為 0.0568。

本節的實驗方法為挑選另外 50 句化學機械研磨的步驟當作測試組，用本系統已訓練的參數來排列出相似序列和訓練過門檻值 0.0568 的 BLEU 排列出相似序列再進行與專家標示的正確序列進行編輯距離計算。本系統方法的平均編輯距離為 1.24，而 BLEU 的平均編輯距離為 3。實驗結果如表格 3 所示也可看出在準確率方面本系統平均準確率 0.952 高於 BLEU 的 0.727，在召回率方面本系統的平均召回率 0.944 亦高於 BLEU 的 0.408。由此實驗可看出本系統提供的方法較自然語言常採用的 BLEU 更適合應用在專利步驟的相似比對上。

表格 3 本系統 VS BLEU 測試五十句的準確率和召回

率

系統方法	平均準確率	平均召回率
本系統	0.952	0.944
BLEU	0.727	0.408

四、結論

本研究的專利文件來源為美國專利商標局(USPTO)[1]。只做方法(method)類的專利文件，其專利的內容描述主要為一個方法的步驟陳述。本系統提出的方法僅分析專利文件的申請範圍(claims)，雖無限定特定領域，目前僅以化學機械研磨領域為案例測試。故本系統企圖解決[4][5]中尚未處理的方法專利，能節省解讀專利申請範圍的人力物力。從專利申請範圍中自動擷取出各項元件、元件之間的關係、步驟、步驟之間的關係，而結構化擷取物成樹狀機讀式檔案，可讓程式能依循定義對檔案作處理，進行相似度比對以避免侵權，以及圖形化檔案加速使用者快速了解一篇專利的架構。本研究成果摘要如下：

1.模擬專利侵權比對中的專利範圍自動解讀
2.自動擷取專利申請範圍中的步驟和步驟關係
3.模擬專利侵權比對中的專利範圍自動解讀
4.自動擷取專利申請範圍中的步驟和步驟關係
5.自動建立方法專利的機讀式檔案
6.步驟相似度比對方法

感謝詞

本研究感謝中華民國經濟部學界科專計劃編號 93-EC-17-A-05-S1-030 三年來的補助。

參考文獻

[1] United States Patent Trademark Office http://www.uspto.gov/

[2] 陳裕禎，“如何解讀專利範圍？”，國立清華大學專利實務暨技轉培訓班講義，新竹，2005.

[3] 陳志超，“專利法-理論與實務”，五南圖書出版股份有限公司，2002.

[4] 林士能，“專利文件語意之擷取與比對”，國立清華大學資訊工程學系，碩士論文，2005.

[5] Chi-Feng Lee, “Automatic Acquisition of Domain Specific Regular Expressions from Patent Documents, Master thesis, NTHU, 2006.

[6] Kristina Toutanova and Christopher D. Manning, “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger”, Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), Hong Kong.

[7] Dan Klein and Christopher D. Manning, “Accurate Unlexicalized Parsing”, Proceedings of the 41st Meeting of the Association for Computational Linguistics, 2003.

[8] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning, “Generating Typed Dependency Parses from Phrase Structure Parses”, In Proceedings of the LREC Conference. Genoa, Italy. 2006

[9] George A. Miller, WordNet: A Lexical Database for English”, Communications of the ACM, Vol.38 No.11, 1995

[10] 王世仁，專利工程導論，俊傑書局，2002

[11] Holland, J., 1975, “Adaptation in Natural and Artificial Systems”, University of Michigan Press, Ann Arbor, MI.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),Philadelphia, July 2002, pp. 311-318.