# 一種主要使用語料庫標記進行歧義校正的、最大匹配漢語自動分詞算法設計

黎邦洋 簡葆 孫朝奮 孫茂松

香港城市理工學院應用語言學系

## 摘要

漢語自動分詞是中文信息處理中特有的一個困難問題。大陸在這方面所做的工作通常採用最大匹配法加規則歧義校正模式，正切率98％左右；海外以詞頻爲基點的統計計算模型的正切率爲95％。本文介紹的分詞算法設計主要採用最大匹配法，但除了語法規則之外，主要使用了語料庫語言學的標記(tag) 並輔以詞頻信息進行概率歧義校正，以期在最大匹配法的基礎上取得進一步的成績。

## 1. 前言

漢語自動分詞是中文信息處理中特有的一個困難問題。八十年代初以來，大陸及海外陸續開展了這方面的研究。迄今爲止，大陸提出的分詞方法達12種以上，通常採用“機械匹配（最大匹配法，記作MM）＋規則歧義校正”模式，正切率在98％左右（正切率定義爲一段漢語語料中，切分正確的詞所含漢字個數與該語料所含漢字總個數之比），且研究規模宏大，最具代表性的分詞系統之一—— CDWS 較成功地處理漢字語料兩千餘萬 [1]～[10]；海外則以詞頻爲基點構造分詞的統計計算模型，思路明顯區別於前者，但一般規模不大，正切率爲95％ [14][15] 。

評價一個分詞系統性能高低的主要技術指標爲正切率，而正切率又直接受歧義切分字段制約。漢語的歧義切分字段包括兩種基本類型：

- 交集型歧義切分字段

定義：在字段 S = $a_1$ ... $a_i b_1$ ... $b_j c_1$ ... $c_k$ 中，若$a_1$ ... $a_i$ $b_1$ ... $b_j$和$b_1$ ... $b_j c_1$ ... $c_k$ 分別構成詞，則字段 S爲交集型歧義切分字段，其中$b_1$ ... $b_j$稱爲交段。

這是常見類型，約占全部歧義切分字段的84％ [10] 。

● 包孕型歧義切分字段

定義：在字段 S =$a_1$ ... $a_i b_1$ ... $b_j$中，若$a_1$ ... $a_i$，$b_1$ ... $b_j$和S均構成詞，則字段 S稱爲包孕型歧義切分字段。

約占全部歧義切分字段的26％ [10] 。

歧義切分字段亦可是這兩種基本類型的組合。

前述兩大類方法各具所長，但也分別存在一些問題：

● 第一大類方法（大陸）

MM法簡單快速（給出的可能切分唯一，此亦即最終的切分結果），且符合人類語言心理。文獻[2] 表明，即使沒有其它任何知識(詞法、句法、語義等)，單獨運用MM法已能滿足一般精度的要求（正切率在90％左右)。其不足是：(1) 無條件地機械運用MM法，某些情況下會屏蔽掉正確切分，特別是無法檢測出包孕型歧義切分字段；(2) 歧義校正部分未定量利用對漢語分詞無疑極其重要的詞頻信息，且校正規則所涉及的句法、語義知識過於零散，不夠系統。由於未引入這些知識，或引入得不夠，使得算法對解決歧義切分字段手段有限，分詞精度難以期望有新的突破。

● 第二大類方法（海外）

先找出所有的可能切分解，然后主要根據詞頻信息進行計算，將可能解空間約束到一個算法認爲的最優解。這類實驗揭示了詞頻在分詞中的重要位置。但 (1) 沒有採用MM法，導致可能解空間膨脹太大（組合爆炸)，造成了許多不必要的干擾；(2) 僅進行單純的統計計算，未爲分詞系統提供進一步的句法、語義分析機制。

## 2. 算法設計

基於對已有分詞方法的認識，我們構造了一種“主要使用語料庫標記

進行歧義校正的、有條件最大匹配漢語自動分詞算法”。總體思路是:(1) MM法作爲一種切實、高效的分詞方法已爲大規模語言工程實踐所肯定，故本算法亦以其作爲基本構架；然而其運用應在一定限制下進行，以避免丟失正確切分；(2) 漢語的複雜性要求分詞系統所依賴的知識必須是多元的，統計知識和規則形式給出的知識從不同角度描寫了語言事實，兩者應盡可能地相互協調、補充。我們的算法按運作順序分爲“切分”和“約束”兩大階段（“約束”相當於歧義校正）。

## 2.1 切分階段

有條件運用MM法，給出此意義下的全部可能切分解。在保證正確解一定存在於可能解空間的前提下，盡可能壓縮後者，爲約束階段降低噪聲干擾。這個過程對待切分字串(通常以句子爲單位)從左向右進行一次掃描。

爲表述方便，首先定義幾個符號和函數:

● 絕對切分符 '|'與可能切分符 'Φ'

設待切分字串 S由 n個漢字組成:

$$S = \underset{p_0}{\ } c_1 \underset{p_1}{\ } c_2 \underset{p_2}{\ } \cdots \underset{p_{i-1}}{\ } c_i \underset{p_i}{\ } c_{i+1} \underset{p_{i+1}}{\ } \cdots \underset{p_{n-1}}{\ } c_n \underset{p_n}{\ }$$

下角標表示位置，位置$p_{i-1}$和$p_i$之間爲漢字$c_i$。 S的結束位置$p_n$亦記作$p_\infty$。

任給 $1 \leq i \leq n$，若形如:

$$c_1 c_2 \cdots c_i \quad c_{i+1} \cdots c_n \qquad (*)$$

出現在S的正確切分解中，則位置$p_i$上存在一個絕對切分符 '|'，即:

$$c_1 c_2 \cdots c_i \mid c_{i+1} \cdots c_n$$

若 (*)出現於 S的可能切分解中，則位置$p_i$上應插入一個相對切分符'Φ'，即有:

$$c_1 c_2 \cdots c_i \ \Phi \ c_{i+1} \cdots c_n$$

顯然，位置$p_i$上的'Φ'只是該位置存在 '|'的必要條件。

● 函數 PMM ($p_s$, $p_e$, $p_m$, w)　　$0 \leq p_s \leq p_e \leq p_m \leq p_\infty$

在 S的某個起始位置$p_s$至某個終止位置$p_e$之間運用MM法得到相應的詞

w 及該詞最後一個字右側位置$p_m$。若 w不存在，則返回 NIL。

● 函數 INS ($p_x$, chr)

在位置$p_x$上插入一個字符chr。兩個常見的操作爲 INS($p_x$, |) 及 INS($p_x$, Φ)。

整個切分過程自始至終由以下四條元規則控制、引導：

元規則1 (MR1)

若 PMM($p_x$, $p_\infty$, $p_m$, $w_{xm}$) 則 INS($p_m$, Φ)。且在下述兩種情形下，必有'Φ'提升爲 '|'：

(1) $w_{xm}$ 具備屬性fc; 或

(2) 元規則2～4不成立

元規則2～4對具備屬性fc的詞不起作用，即這類詞擁有最高的切分優先權 (如成語、熟語及專業詞彙等)。

三里島| 的| 壓水堆| 的| 事故| 與| 切爾諾貝爾| 核電站| 事故| 的| 嚴重|程度| 是| 相近| 的| 。| (例1)

元規則2 (MR2)

若 PMM($p_x$, $p_\infty$, $p_m$, $w_{xm}$) 且

對某個位置$p_y$ ($p_x \leqq p_y \leqq p_m$) 有

PMM($p_y$, $p_m$, $p_z$, $w_{yz}$) 且 $w_{yz}$ 具備屬性fΦ

則除應用元規則1 外，尚有 INS($p_y$, Φ)及 INS($p_z$, Φ)。

(a) 他馬上就來。

(b) 他從馬上下來。 (例2)

若僅對(b)施以元規則1，則會丟掉正確切分 (這裡，“馬上”對應元規則2中的$w_{xm}$，“上”對應$w_{yz}$)。這類操作是由諸如“上”之類的詞的特點所決定的。此屬性記作fΦ，具fΦ屬性的詞在漢語中爲封閉集，常對應某些詞類，如方位詞、量詞、介詞、副詞等，且多爲單字詞。

元規則3 (MR3)

若 PMM($p_x$, $p_\infty$, $p_m$, $w_{xm}$) 且 PMM($p_{m-1}$, $p_\infty$, $p_n$, $w_{(m-1)n}$) 且
PMM($p_x$, $p_{m-1}$, $p_z$, $w_{xz}$) 且 $p_{m-1} = p_z$
則除應用元規則1 外，尚有 INS($p_z$, Φ)。

(a) 研究生應該努力鑽研。
(b) 研究生命起源。 (例3)
(b) 滿足元規則3，則必須保持可能切分“研究Φ生命...”。

元規則4 (MR4)

若 PMM($p_x$, $p_\infty$, $p_m$, $w_{xm}$) 且 PMM($p_m$, $p_\infty$, $p_n$, $w_{mn}$) 且
PMM($p_x$, $p_{m-1}$, $p_k$, $w_{xk}$) 且 PMM($p_k$, $p_m$, $p_l$, $w_{kl}$) 且
$p_m = p_l$ 且 $w_{mn}$ 與 $w_{kl}$ 之間存在某種句法語義制約關係
則除應用元規則1 外，尚有 INS($p_k$, Φ)。

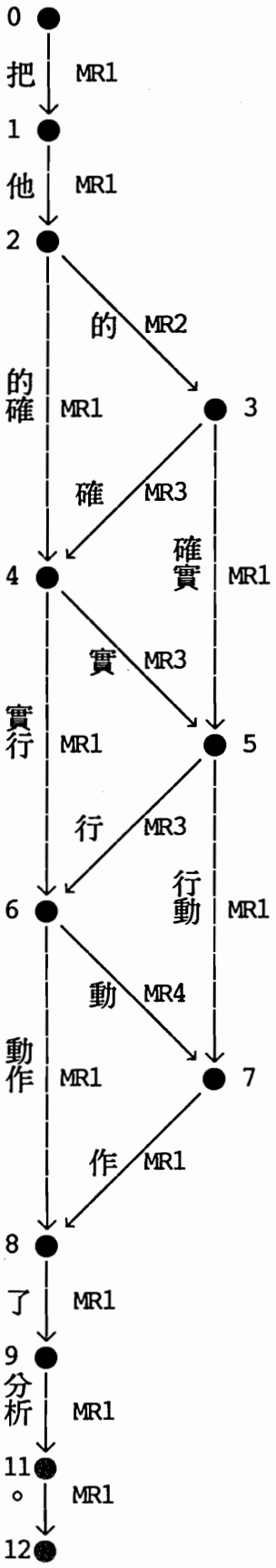
(a) 機器翻譯屬於計算語言學領域。
(b) 這句話，機器翻譯起來很難。 (例4)

(b) 中，“翻譯”爲動詞，“起來”爲趨向動詞，而模式“動詞 +趨向動詞”在句法上結合的可能性很大，故可能切分“機器Φ翻譯起來...”應予保留。而 (a)無類似(b) 的語境，僅可運用元規則1 。

元規則3、4利用了一條重要假設，其可靠性已得到大量實驗的充分支持 [2]：
交集型歧義切分字段的交段長度爲1。

圖1 給出文獻[14]所舉之例句經切分階段處理後得到的狀態：
把他的確實行動作了說明。 (例5)

切分階段應貫穿的原則是：
(1) 潛在的可能性一定保留全；
(2) 不可能的切分盡早予以排除(剪枝)；




把 他 的 確 實 行 動 作 了 分 析 。
0 1 2 3 4 5 6 7 8 9 10 11 12

切分階段給出的分析結果如圖左所示。

該圖可展開爲13條可能切分路徑：

(1) 把 他 的確 實行 動作 了 分 析 。
(2) 把 他 的確 實行 動 作 了 分 析 。
(3) 把 他 的確 實 行 動作 了 分 析 。
(4) 把 他 的確 實 行 動 作 了 分 析 。
(5) 把 他 的確 實 行動 作 了 分 析 。
(6) 把 他 的 確 實行 動作 了 分 析 。
(7) 把 他 的 確 實行 動 作 了 分 析 。
(8) 把 他 的 確 實 行 動作 了 分 析 。
(9) 把 他 的 確 實 行 動 作 了 分 析 。
(10) 把 他 的 確 實 行動 作 了 分 析 。
(11) 把 他 的 確實 行 動作 了 分 析 。
(12) 把 他 的 確實 行 動 作 了 分 析 。
(13) 把 他 的 確實 行動 作 了 分 析 。

圖1

(3) 所使用的詞法、句法、語義知識以規則形式寫出，不追求完備性，但一定要可靠，亦應注意其一般性。

## 2.2 約束階段

切分階段得到一組可能切分解，約束階段則要找出其中的正確解（不一定唯一），實質上就是對歧義切分字段的處理。文獻[2]指出，歧義切分字段的長度一般≦6 。這是一個很好的性質，可把歧義切分字段的影響限制在一定範圍內而不致傳播、蔓延。

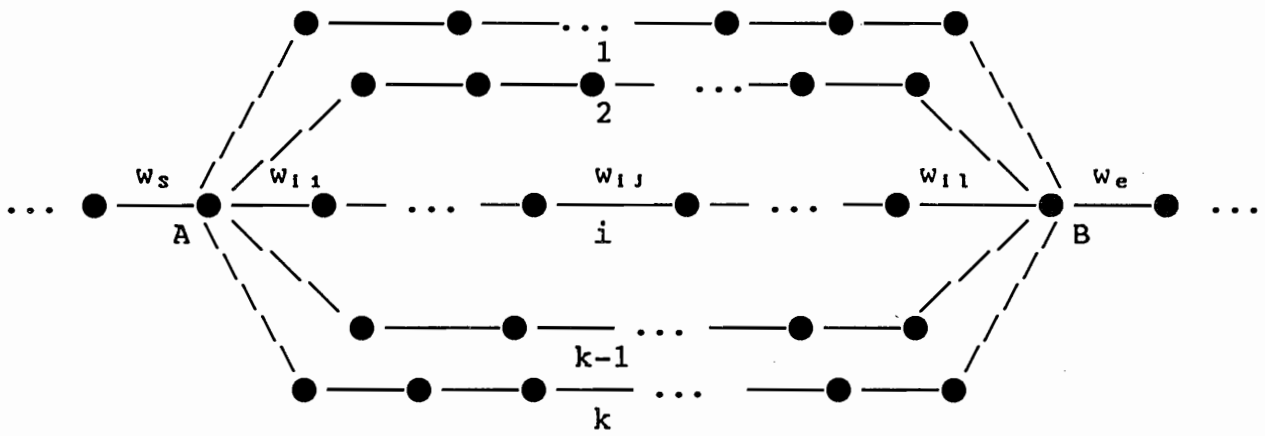



圖2

圖2中，位置 A與 B之間爲歧義切分字段。該字段存在k種可能切分，分別對應從A 到 B的k條路徑（圖1中，位置2與8之間的可能切分路徑數目k=13）。第i條路徑切分階段則給出的詞序列爲：

$$w_{i1}\ldots w_{ij}\ \ldots\ w_{il}\quad (i=1,\ldots,k,\ \sum_{j=1}^{l} \text{length}(w_{ij}) \leqq 6),$$

詞$w_s$, $w_e$鄰接$w_{i1}$, $w_{il}$，且切分無歧義。則約束階段則相當於從 k條路徑中選擇分詞意義的最佳路徑。

解決歧義切分字段最“自然”不過的對策是訴諸句法、語義規則，進一步地，當遇到歧義切分點時，即調用Parser來判斷，如果分析失敗，再Feedback回來。這種分詞與Parser一體化的方法似乎非常之徹底，但從實用的觀點看，並不十分可行。主要會受到三個問題的困擾：(1) 囿於目前

漢語Parser的研究水平及其能力（即使分詞完全正確，也只能在某些領域實現有限程度的句法、語義分析），任何期望倚重Parser作爲解決歧義切分字段之手段的想法似尚缺乏現實的基礎；(2) 這種狀況導致了分詞系統所能利用的句法、語義規則必然是局部的，基本上僅涉及若干毗鄰詞之間的線性關係，並不充分、可靠（沒有反映句子中各成分之間的層次關係，故而還不能算作眞正意義的句法、語義分析)； (3) 由於(1)和(2)，分詞系統無法建立完整的規則體系，往往陷於“就事論事”而難以自拔，也難以進行規則之間的無矛盾性檢查，並且規則的置信度不易定量描述。我們的算法則借鑒了語料庫語言學的思想。

語料庫語言學 (Corpus Linguistics) 80年代才嶄露頭角，其最引人注目的成果之一是對庫存語料進行語法標注 (grammatical tagging)。1970-1978 年間，Brown 大學實施了對 BROWN庫的標注，共採用86種詞類標記(TAG)，設計了基於規則的自動標注系統TAGGIT（擁有3300條規則），對100萬詞語料自動標注的正確率爲77％。1978-1983年間Lancaster、Oslo、Bergen三所大學聯合對LOB庫開展類似的標注實驗，設計的CLAWS系統採用133種詞類標記。意義重大的是，他們完全放棄了 TAGGIT系統那種傳統的規則模型，而把自動標注的算法建立在統計信息的基礎上：首先利用已帶有語法標記的語料獲取兩個相鄰標記的同現頻率，據此建立一個133×133的“標記轉移頻率矩陣”，整個標注過程所依據的“知識”均由此矩陣提供。CLAWS 對 100萬詞的 LOB庫實行語法標注的正確率已達96％，比基於規則的TAGGIT提高了將近20％，有效地處理了同形和兼類問題[13][17]。

我們認爲，漢語分詞中的歧義切分字段處理問題與語料庫自動標注問題具有可類比性，因此在方法上應基本是相通的。約束階段中運用關於標記的統計計算至少可有幾點益處：(1) 實際上將句法、語義知識融於統計數據中，表示簡潔，避免了採用規則系統可能導致的種種問題，且據之構造的算法可在整個約束過程中貫徹到底；(2) 統計數據直接從不受任何限制的實際語料中獲得，可有效提高分詞系統的能力及覆蓋面，並且分詞結果能隨時反饋到統計數據中，不斷求精，使系統有一定的自學習功能；(3) 標記數目通常不多，對應的“標記轉移頻率矩陣”所需空間普通PC機即可支持，利於系統的推行、應用。

我們初步擬定了一組面向分詞的句法、語義標記，主要以句法範疇爲主，兼顧語義分類（文獻[10]指出，近95％的歧義切分字段可用詞法、句法知識解決，僅3.5％及1.7％左右才需分別訴諸語義、語用知識)， 並將在今後的實驗中不斷修正。實驗系統處理的語料爲香港城市理工學院另一

研究項目ACTAS 所涉及的關於大亞灣核電站安全問題的40篇社論，約 8萬字 [21]。

設標記集

$TAGSET = \{ tag_i \mid i = 1, \ldots, n \}$

則概率 $p(tag_i)$

及條件概率 $p(tag_j \mid tag_i)$ $(i,j = 1, \ldots, n, \ i \neq j)$

可從這個語料中（預先進行人工標記）計算而來。

參閱圖2 。設$tag_s$, $tag_{i1}$, ... , $tag_{i1}$, $tag_e$分別爲對應$w_s$, $w_{i1}$, ... , $w_{i1}$, $w_e$的標記，則$tag_s$,$tag_{i1}$, ... ,$tag_{i1}$, $tag_e$的同現概率爲：

$$p(tag_s tag_{i1} \ldots tag_{i1} tag_e)$$
$$= p(tag_s \mid tag_{i1} \ldots tag_{i1} tag_e)\ p(tag_{i1} \mid tag_{i2} \ldots tag_{i1} tag_e)$$
$$\ldots$$
$$p(tag_{i1} \mid tag_e)\ p(tag_e)$$

上式中的每一項可參考文獻 [18][19][20] 設計適當算法轉化爲對有關 $p(tag_i)$及$p(tag_j \mid tag_i)$ 的一組計算求得。

對 k條路徑依上式計算，取得最大值的路徑爲最佳解。

實際運作中，會遇到一個詞下掛有多個標記的情形（兼類詞或者多義詞），只需將其再擴展爲若干路徑即可，仍如上處理。此時，一旦分詞完成，句子的兼類詞也就基本解決了（多義詞僅能在標記集所涉及的語義層面上部分解決）。

標記集選取是否合理直接影響歧義切分字段的校正精度。此外，標記集終究只是對句法、語義範疇的一個近似概括，這個特點決定了必存在此種方法“鞭長莫及”的死角。故本算法還引入了兩種輔助手段：

● 輔助手段之一 —— 句法語義規則 (AM1)

我將來上海。 (例6)

切分過程給出兩種可能切分：

(a) 我 |將來 Φ上海 |。|

(b) 我 |將 Φ來 Φ上海 |。|

設當前指針指向“來”，則有規則：

如果全句無其它動詞且當前指針所指詞的左鄰詞爲時態副詞，取(b)

約束用句法語義規則按其功用大致分三種類型。第1類：可靠的(如詞法規則與處理例6的規則)；第2類：針對標記集法的弱點專門設計的(如一些與連詞有關的現象)；第3類：雖不那麼可靠但仍有一定參考意義。其中第1、2類規則多與具體的詞直接發生關聯，即多爲個性規則 (individual rule)。

● 輔助手段之二 — 基於詞頻的計算 (AM2)

對圖2，此意義下的最佳解爲：

$$\text{最佳解} = \{ \text{第 } t\text{條路徑} \mid t = \max_{i=1}^{k} p(w_s)p(w_e)\prod_{j=1}^{l} p(w_{ij}) \}$$

實際上將 $w_s$, $w_{i1}$, ... , $w_{il}$, $w_e$ 視作獨立變元處理 (零階Markov鏈)，故可能導致錯誤。

TAG法與AM1, AM2整合之過程爲：

(1) 調用 AM1中第1、2類約束用句法語義規則。
(2) 若步驟(1)能得到唯一切分，則輸出之，約束階段結束；否則轉步驟(3)。
(3) 對所有可能切分進行基於TAG的路徑計算，其值按降序排列爲：
    $\{val_1, val_2, val_3, \ldots\}$
    若 $|val_1 - val_2| <= t0$ (t0爲一threshold值，由實驗測定)
    則轉步驟(4)；
    否則，對應值$val_1$的路徑爲正確解，輸出之，約束階段結束。
(4) 調用AM1中第3類約束用句法語義規則及進行AM2計算。根據兩者各自給出的 preference及其強弱，最終決定取對應值$val_1$抑或$val_2$的路徑爲正確解，還是兩條路徑均可。至此，約束階段全部結束。

## 3. 幾個正在研究的相關問題

圍繞本算法的實現，我們正在研究的相關問題(部分已取得結果)主要

有：

(1) 漢語“詞”的定義及判斷標準；
(2) 支持本算法的機器用分詞詞典；
(3) 切分元規則2 所涉及的具有屬性fΦ的詞及其與詞類的關係；
(4) 切分元規則4 所涉及的常見句法、語義制約類型；
(5) 面向分詞的漢語句法、語義標記集；
(6) 簡單的規則描述語言及規則解釋程序。

整個程式將在 286機上用 C語言實現，估計可於九一年底完成。

## 主要參考文獻


[1] 劉源，梁南元，“漢語處理的基礎工程 — 現代詞頻統計”，中文信息學報，1986年，第1期
[2] 梁南元，“書面漢語自動分詞系統 — CDWS”，中文信息學報，1987年，第2期
[3] 揭春雨，劉源，梁南元，“論漢語自動分詞方法”，中文信息學報，1989年，第3期
[4] 梁南元，“漢語計算機自動分詞知識”，中文信息學報，1990年，第2期
[5] 周依欣，吳蔚天，“一種實用的漢語切分方法 ... 鏈接表法”，中文信息學報，1990年，第2期
[6] 楊抒，伊波，“基於後加詞典利用句法語義知識的漢語詞切分檢糾錯方法”，中文信息，1989年
[7] 姚天順，張桂平，吳映明，“基於規則的漢語自動分詞系統”，中文信息學報，1990年，第1期
[8] 黃祥喜，“書面漢語自動分詞的生成 — 測試方法”，中文信息學報，1989年，第4期
[9] 李國臣，劉開瑛，張永奎，“漢語自動分詞及歧義組合結構的處理”，中文信息學報，1988年，第3期
[10] 何克抗，徐輝，孫波，“書面漢語自動分詞專家系統總體設計報告”北京師範大學現代教育技術研究所，1990年11月
[11] 陶沙，“高頻詞”，中文信息，1986年

[12] 劉湧泉，“談談詞庫問題”，中文信息學報，1986年，第1期
[13] 黃昌寧，“語料庫語言學”，中國計算機用戶，1990年，第11期
[14] Charng-Kang Fan, Wen-Hsiang Tsai, Automatic Word Identification in Chinese Sentences by the Relaxation Technique, Computer Processing of Chinese & Oriental Languages, Vol.4, No.1, November 1988, pp33-56
[15] Richard Sproat, Chilin Shih, A Statistical Method for Finding Word Boundaries in Chinese Text, Computer Processing of Chinese & Oriental Languages, Vol.4, No.4, March 1990
[16] K.T. Lua, From Character to Word — An Application of Information Theory, Computer Processing of Chinese & Oriental Languages, Vol.4, No.4, March 1990
[17] Roger Garside, Geoffrey Leech, Geoffrey Sampson, The Computational Analysis of English — a Corpus-based Approach. Longman Group UK Limited, 1987
[18] Shachter, R.D., Intelligent Probabilistic Inference, in: Kanal, L.N. and Lemmer, J.F. (eds.), Uncertainty in Artificial Intelligence, (North-Holland, Amsterdam, 1986) pp371-382
[19] Pearl, J., Fusion, Propagation and Structuring in Belief Networks, Artificial Intelligence, 29(1986) pp241-288
[20] David J. Spiegelhalter, Probabilistic Reasoning in Predictive Expert Systems, in: Kanal, L.N. and Lemmer, J.F. (eds.), Uncertainty in Artificial Intelligence (North-Holland, Amsterdam, 1986)
[21] B K T'sou, H C Ho, H L Lin, Lun Suen Caesar, G K F Liu, A Y L Heung, Automated Chinese Text Abstraction: A Human-machine Co-operative Approach, Computer Processing of Chinese & Oriental Languages, Vol.5, No.1, September 1990, pp33-42