Indexing and Abstracting

Lecture 08 – Text Processing

Kuang-hua Chen
Department of Library and Information Science
National Taiwan University

背景

- 對於生活於20世紀的人類而言，世界的變化是急遽又富挑戰性的
- 各種訊息傳遞技術的發明，使得資訊的傳遞速率越來越快
- 通訊與電腦技術的結合，網際網路的全球連線，更將人類的生活方式帶進前所未有的境地
- 索引與檢索不可避免地也進入的網路世界

背景（續）

- 透過網路連結各大圖書館，待在家裡就能夠查閱圖書館的館藏，一旦確認有需要的圖書，亦可線上預約
- 若有電子版本，則更可以線上覽讀，將電子圖書館視爲家中書房的延伸，足不出戶就能飽讀群書

網際網路

- 網際網路可說是世紀末超級的新興媒體，原本僅是學術上用於溝通訊息的通訊管道，經由WWW迷人的瀏覽介面，以新的面貌讓網路使用者更方便地使用網際網路
- 這種新型態的資訊載體展現超乎想像的功能，也讓普羅大眾體會了看不到、摸不到卻又無所不在的真實感受

網際網路（續）

- 網際網路已經將媒體的掌握權解放了，任何人都可以將個人的意見、出版品放到網路上，不再受制於出版商
- 實際使用網際網路幾年後，更能夠掌握它的特性
- 各種型態的資訊陸續出現，各種格式的資料推陳出新，網際網路呈現百家齊鳴的現象

網際網路（續）

- 網際網路持續地蓬勃發展，透過網路傳輸的資訊量越來越大
- 使用者以往苦惱於無法取得資訊，今日卻面臨了資訊爆炸的問題
- 獲得有用資訊的代價越來越高

資訊服務

- 如何協助讀者或是尋求資訊的人們取得有用的資訊，成爲圖書館學與資訊科學研究領域中非常重要的課題
- 資訊檢索系統早期在獨立而封閉的環境運作，今日則處於開放的環境
- 無論面對的環境如何改變，其協助讀者或使用者取得有用、適用資訊的目標卻無二至

網際網路的原生性服務

- 提供吾人取得資訊的基本功能
- 遠端登錄（Telnet）
- 電子郵件（E-mail）
- 新聞討論群（Usenet）
- 全球資訊網（World Wide Web）

WWW

- WWW在1993年Mosaic瀏覽器推出後，迅速吸引學術界的眼光
- 經由Netscape公司與Microsoft公司在瀏覽器市場的爭霸戰
- 商業體系推波助瀾，WWW的使用者迅速攀升

網際網路的加值型服務

- 搜尋引擎（Search Engine）
  - 使用者可以取得特定事件的相關資訊
- 主題指引（Subject Directory）
  - 透過主題指引，使用者可以取得相關主題的資訊
- 其他服務
  - BigFoot、Four11等找人的服務
  - 搜尋e-mail的服務

搜尋引擎

國　外	國　內
AltaVista (http://altavista.digital.com/)	GAIS（蓋世） (http://gais.cs.ccu.edu.tw/)
Lycos (http://www.lycos.com/)	WhatSite（哇塞） (http://www.whatsite.com/)
OpenText (http://www.opentext.com/)	聚寶盆 (http://spring.nii.nchc.gov.tw/Search/)
Northern Light (http://www.northernlight.com)	Openfind (http://www.openfind.com.tw/)
InfoSeek (http://info.infoseek.com/)	
Excite (http://www.excite.com/)	
HotBot (http://www.hotbot.com/)	
Google (http://www.google.com/)	
WebCrawler (http://webcrawler.com/)	

主題指引

國　外	國　內
Yahoo (http://www.yahoo.com/)	Yam（蕃薯藤） (http://taiwan.ntu.edu.tw/)
Galaxy (http://www.galaxy.com/)	
PlanetSearch (http://www.planetsearch.com/)	
StartPoint (http://www.stpt.com/)	
The WWW Virtual Library (http://vlib.stanford.edu/Overview.html)	
Magellan (http://www.mckinley.com/)	
Deja News (http://www.dejanews.com/)	

問題

- 雖然眾多搜尋引擎與主題指引已經提供吾人相當大的幫助，並且提供簡短的文件描述
- 檢索所得的文件仍然相當的多，而且簡短的描述通常無法判斷該文件是否爲相關的文件
- 使用者必須連結檢索所得的文件，真正閱讀之後才能夠知道文件是否適用

影響

- 吾人並沒有真正享受上述服務帶來的好處，很可能隨著文件的來來往往，卻沒有需要的文件，心情越來越沮喪
- 文件的來來往往，使得網路流量大增，卻沒有達到實際的效用，造成網路不必要的負擔

訊息性的資訊服務

- 在進入21世紀的關鍵時期，在NII、GII、NGI等口號震天價響的新資訊時代，享用更好的資訊服務，並非是過分的要求
- 是否有啓發性的資訊服務讓吾人更有效地取得所需的資訊？
- 重要的研究目標
  - 資訊擷取（Information Extraction）
  - 自動摘要（Automatic Summarization）
  - 資訊過濾（Information Filtering）
  - 主題偵測與追蹤（Topic Detection and Tracking）

Human

- 人的長處
  - 優雅
  - 彈性
- 人的短處
  - 不一致
  - 曖昧
  - 倦怠

Hand-Coded Processing

- Human knowledge
- Human value
- Error-prone
- Effort-consuming
- Inconsistency

Automatic Processing

- Cost-saving
- Time-saving
- Consistency
- Virtual Reality

Compromise

- Semi-automatic processing
- Bootstrapping

Applications

- Automatic Indexing
- Automatic Clustering
- Automatic Summarization
- Information Retrieval
- Information Extraction
- Question and Answer (Q&A)

Models

- Model human behavior
- Model human mind
- Artificial Intelligence
- Examples
  - how did you select index terms?
  - how did you make summary?
  - how did you design models to fulfill human tasks?
  - how did you know the models is appropriate for the designated tasks?

Transforming Environment

- Standalone systems
- Organization-based systems
- Internet-based systems
- Intranet-based systems

Transforming Techniques

- Language techniques
  - Part-of-speech tagging
  - Parsing
  - WSD
  - MT
  - Discourse analysis
- IR techniques
  - Boolean model
  - Vector-space model
  - Probabilistic model

Techniques for Text Processing

- Researches of natural language processing (NLP) have developed many high-performance analysis systems.
- Tokenization
- Stemming and Tagging
- Sense tagging
- Syntactic analysis (parsing)

Tokenization

- The performance of tokenization module is about 98% correct rate [Palmer and Hearst, 1994].
  - The difficulty of this part is to distinguish whether periods are full-stop or part of abbreviations.

Stemming and Tagging

- The Stemming module is also good enough.
  - Porter algorithm [Porter, 1980]
  - Two-level morphology [Karttunen, 1983].
- Lexical analysis module, the most successful part of researches of NLP in recent years.
  - Probabilistic tagger [Church, 1988]
  - Rule-based tagger [Brill, 1994]
  - Hybrid tagger [Voutilainen, 1993]
  - Finite-state tagger [Kempe, 1997]

Word Segmentation

- Chinese word segmentation
  - 將黃大目的確實行動作了解釋
    （改寫自張俊盛教授舉的例子）
  - 將◆黃大目◆的◆確實◆行動◆作◆了◆解釋
- Segmentation tools
  - 中央研究院詞庫小組
  - 致遠科技
  - 清華大學自然語言處理實驗室
  - 臺灣大學自然語言處理實驗室
  - 工研院電通所

Approaches of Word Segmentation

- The longest word first
- The probabilistic-based
- Integrating tagging and segmentation
- Take proper nouns into consideration

Syntactic Analysis

- A challenging work
- From the viewpoint of NLP, the correct and complete parse tree is very important
- For applications like IR and IE, time is the most critical factor
- Leverage of time and correctness is important
- Partial parsing

Partial Parsing

- Fidditch [Hindle, 1983]
- Chunker
  - Rule-based chunker [Abney, 1991]
    - [The effort] [to establish] [such a conclusion] [of course] [will have] [two foci].
  - Probabilistic chunker [Chen and Chen, 1993]
    - [When we] [are about to] [read a sentence,] [we usually read it] [chunk by chunk].
    - [楊貴妃] [進入] [後宮] [前]

Partial Parsing (continued)

- Transformational-based parser [Brill and Marcus, 1992]
  - [[The_AT [daring_JJ boy_NN]] [[chased_VBD] [him_PPO]]].
- Statistical binary parser [Chen, 1998]
  - [[[[Jack_NP Young_NP] [is_BEZ also_RB]] [[a_AT doubtful_JJ] starter_NN]] [next_AP year_NN]]
- Finite-state parser

Word Sense Tagging

- Rule-based [Chen and Chen, 1991]
- Knowledge-based [Ward, 1988; Baker *et al.*, 1994]
- Dictionary-based [Lesk, 1986]
- Example-based [Ng and Lee, 1996]
- Statistical Approach
  - Corpus-based [Brown *et al*, 1991]
  - Dictionary-based [Guthrie *et al.*, 1992; Niwa and Nitta, 1994]
  - Corpus and dictionary-based [Doi and Muraki, 1993]
- Neural Network Approach [Veronis and Ide, 1990]

Research Trends

- Construct benchmark collection
  - TREC (Text REtrieval Conference)
  - NTCIR (NACSIS Test Collection for IR Systems, now NII Test Collection for IR Systems)
  - AMARYLLIS
  - CTREC (Chinese Text REtrieval Conference)
  - MIRA (Evaluation Frameworks for Interactive Multimedia Information Retrieval Application)
  - CIRB (Chinese Information Retrieval Benchmark)
  - HANTEC (Korean Test Collection)
- Researches on controlled-vocabulary indexing
- Researches on machine-aided automatic indexing
- Researches on phrase term indexing (pre-coordinated)

Research Trends (continued)

- Researches on PAT tree
- Researches on thesaurus construction
- Researches on parsing techniques
- Researches on discourse Analysis (text modeling)