國立中山大學九十二學年度博士班招生考試試題

科目：資訊管理論文評述第一節【資管系】

共1頁第1頁

請您任選下列一個您較熟悉的主題，並根據這個主題設計出一個資管的研究計畫書：

1. 影響員工使用知識管理系統因素之研究
2. 影響電子商務網站顧客關係與滿意度因素之研究
3. 資管人員與使用者衝突及政治鬥爭之研究

計畫書內容請包括

- 研究方法的選擇與說明：質化研究或量化研究。
- 量化研究方面，請包含研究架構（理論基礎、研究模式）、研究變數、研究假說、研究方法與步驟（如抽樣方法、統計分析方法）的描述及說明。
- 質化研究方面，則請包括研究策略的選擇、資料蒐集方法、資料分析方法等等。
- 如果有任何其它您覺得應該涵蓋的內容，也請包括在內。

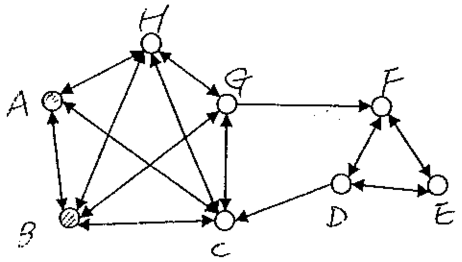
（50%）

國立中山大學九十二學年度博士班招生考試試題

科目：資訊科技論文評述第一節【資管系】 共 / 頁 第 / 頁

Read the attached paper and answer the following questions. For your convenience, the last page of the attachment gives a brief description and an example network about the maximum-flow minimum-cut theorem. Note that the time is limited, and you should budget your time carefully. You are suggested to spend 50 minutes in reading the paper and another 50 minutes in answering the questions.

1. Consider the following web graph. Let S={A, B}. Please show the induced graph G and the capacity $c(v, u)$ of each edge assigned by the procedure EXTRACT-MAX-COMMUNITY, and how it decides the community to which S belongs. (20%)



2. Do you find the design of edge capacity assignment reasonable? Why or why not? (15%)
3. This paper mentions about some other work on the construction of web page communities but does not compare the proposed approach with the others. If you are in a position to evaluate these different approaches, how will you do that? What are the performance metrics you will use to measure the different approaches? (15%)

Self-Organization of the Web and Identification of Communities*

Gary William Flake, Steve Lawrence, C. Lee Giles, Frans M. Coetzee
{flake,lawrence,giles,coetzee}@research.nj.nec.com
NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
Phone: +1 609 951 2795 (Flake) Fax: +1 609 951 2488

Abstract

Despite the decentralized and unorganized nature of the web, we show that the web self-organizes such that communities of highly related pages can be efficiently identified based purely on connectivity. This discovery allows the identification of communities independent of, and unbiased by, the specific words used by authors. Applications include improved search engines, content filtering, and objective analysis of relationships within and between communities on the web.

1 Introduction

The existence of an increasing percentage of human knowledge and society in hyperlinked form on the web has advantages beyond the commonly stated improvements to information access. The potential for analysis of interests and relationships within science and society are great. However, analysis of content on the web is difficult due to the decentralized and unorganized nature of the web. Information on the web is authored and made available by millions of different individuals, operating independently, and having a variety of backgrounds, knowledge, goals, and cultures. We show that, despite its decentralized, unorganized, and heterogeneous nature, the web self-organizes such that the link structure allows efficient identification of communities.

Identification of communities on the web is significant for several reasons. Practical applications include automatic web portals and focused search engines, content filtering, and complementing text-based searches. More importantly, global community identification allows for analysis of the entire web and the objective study of relationships within and between communities (for example, scientific disciplines or countries). Such research could provide insight into the organization and interests of sectors of society, which individual members reflect by their linking practices. For example, links between scientific disciplines may allow more timely identification of emerging interdisciplinary connections.

The web can be modeled as a graph where vertices are web pages and hyperlinks are edges. We define a web *community* as a collection of web pages such that each member page has more hyperlinks (in either direction) within the community than outside of the community (this definition may be generalized to identify communities with varying sizes and levels of cohesiveness). Community membership is a function of both a web page's outbound hyperlinks as well as all other hyperlinks on the web; therefore, these communities are "natural" in the sense that they are collectively organized by independently authored pages. We show that the web self-organizes such that these link-based communities identify highly related pages.

In comparison to previous methods of finding related pages on the web (see the sidebar), our method retains the transparency of methods such as co-citation and bibliographic coupling in explaining why pages are members of the community, yet can identify web communities of arbitrary diameter. Our algorithm

*Published as: G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee. Self-Organization and Identification of Web Communities. *IEEE Computer*, 35(3), 66–71, 2002

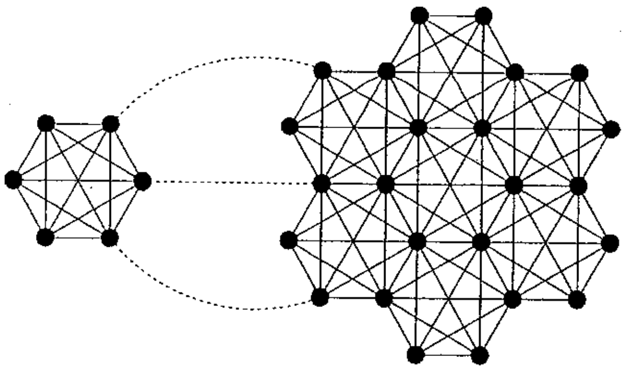


Figure 1: A simple community identification example. Maximum flow methods will separate the two subgraphs with any choice of source vertex s from the left subgraph and sink vertex t from the right subgraph, removing the three dashed links. As formulated with standard flow approaches, all community members must have at least 50% of their links inside of the community; however, additional artificial links can be used to change the threshold from 50% to any other desired threshold. Thus, communities of various sizes and with varying levels of cohesiveness can be identified and studied.

achieves this performance using only link information, without the text information used by algorithms such as HITS. In the absence of full natural language processing, the creation of an explicit link by a web author can be a stronger indication of relevance than implied links generated by the simple phrase and structure matching used by textual methods. In addition, this separation of link structure from content allows us to independently validate the performance of the link-based community estimation with content-based similarity measures.

Identifying a naturally formed community—according to our definition—is intractable in the general case because the basic task maps into a family of NP-complete graph partitioning problems [6]. However, if one assumes the existence of one or more *seed* web sites and exploits systematic regularities of the web graph [3, 8, 10], the problem can be recast into a framework that allows for efficient community identification by using a polynomial time algorithm that should scale well to studying the entire web graph.

2 Maximum Flow Communities

We recast the problem into a maximum flow framework which analyzes the flow between graph vertices. The s-t maximum flow problem [1] is defined as follows. Given a directed graph $G = (V, E)$, with edge capacities $c(u, v) \in \mathbb{Z}^+$, and two vertices, $s, t \in V$, find the maximum flow that can be routed from the source, s, to the sink, t, that obeys all capacity constraints. Intuitively, if edges are water pipes and vertices are pipe junctions, then the maximum flow problem tells you how much water you can move from one junction to another. The Max Flow-Min Cut theorem of Ford and Fulkerson [5] proves that the maximum flow of the network is identical to the minimum cut that separates s and t. Many polynomial time algorithms exist for solving the s-t maximum flow problem [7].

Figure 1 shows the basic intuition of our approach. We choose one or more seed sites to play the role of the source vertex. We require that the sum total of edges connected to the seed sites be greater than the size of the cut set (the dashed edges in Figure 1). If this constraint is not met, then our procedure will only identify a subset of the community with the worst case being that only seed sites will be discovered as being in the community.

One could imagine using an approximate centroid of the web graph (e.g., Yahoo!) as the sink; however,

```
procedure EXACT-FLOW-COMMUNITY
    input: graph: $G = (V, E)$ ; set : $S \subset V$ ; integer : $k$ .
    Create artificial vertices, $s$ and $t$ and add to $V$.
    for all $v \in S$ do
        Add $(s, v)$ to $E$ with $c(s, v) \equiv \infty$.
    end for
    for all $(u, v) \in E$ do
        Set $c(u, v) \equiv k$.
        if $(v, u) \notin E$ then add $(v, u)$ to E with $c(v, u) \equiv k$.
    end for
    for all $v \in V$, $v \notin S \cup \{s, t\}$ do
        Add $(v, t)$ to E with $c(v, t) \equiv 1$.
    end for
    call : MAX-FLOW ($G$ , $s$ , $t$ ).
    output: all $v \in V$ still connected to $s$ .
end procedure
```

```
procedure APPROXIMATE-FLOW-COMMUNITY
    input: set : $S$ .
    while number of iterations is less than desired do
        Set $G = (V, E)$ to fixed depth crawl from $S$.
        Set $k$ to $|S|$.
        call : $C$ = EXACT-FLOW-COMMUNITY ($G, S, k$ ).
        Rank all $v \in C$ by number of edges in $C$.
        Add highest ranked non-seed vertices to $S$.
    end while
    output: all $v \in V$ still connected to $s$ .
end procedure
```

Table 1: Algorithms for identifying web communities. EXACT-FLOW-COMMUNITY augments the web graph in three steps: an artificial source, s, is added with infinite capacity edges routed to all seed vertices in S; each preexisting edge is made bidirectional and rescaled to a constant value k; and all vertices except the source, sink, and seed vertices are routed to the artificial sink with unit capacity. After augmenting the web graph, a residual flow graph is produced by a maximum flow procedure. All vertices accessible from s through non-zero positive edges form the desired result and satisfy our definition of a community. APPROXIMATE-FLOW-COMMUNITY takes a set of seed web sites as input, crawls to a fixed depth including inbound hyperlinks as well as outbound hyperlinks (with inbound hyperlinks found by querying search engines), applies EXACT-FLOW-COMMUNITY to the induced graph from the crawl, ranks the sites in the community by the number of edges each has inside of the community, adds the highest ranked non-seed sites to the seed set, and iterates the procedure. The first iteration may only identify a very small community; however, when new seeds are added, increasingly larger communities can be identified. Note that k is heuristically chosen.

our method works without an explicit sink site via graph augmentation as described in Table 1. See [4] for the corresponding theorem and proof.

If one has access to the entire web graph, then EXACT-FLOW-COMMUNITY will return a set of web pages that obeys our definition of a community because the maximum flow procedure is guaranteed to always find a bottleneck from the source to the sink. Thus, any page that remains connected to the source must have more hyperlinks in the community than outside of the community; otherwise, a more efficient cut would have been to move the web site in question to the non-community.

In EXACT-FLOW-COMMUNITY, the artificial sink is generic in the sense that it is on the receiving end of an edge from every other vertex in the graph. Thus, separating the source from the sink finds a community that is strongly connected internally, but relatively disconnected externally to the rest of the graph.

Table 1 also shows an approximate version of the approach, APPROXIMATE-FLOW-COMMUNITY, which uses a subset of the web graph found by a fixed depth crawl that follows both inbound and outbound hyperlinks. Results are improved on each iteration by reseeding the algorithm with additional web sites found in the earlier steps. Our experimental results were found with the approximate version. However, we also note that the dynamic nature of the web can be exploited with a simpler iterative approximate algorithm that tests for new candidate community members by counting the number of candidate links that fall within the preexisting community.

Francis Crick Community	
Score	**Site Title or Description**
80	*Biography of Francis Harry Compton Crick* (Nobel Foundation)
79	*Biography of James Dewey Watson* (Nobel Foundation)
51	*The Nobel Prize in Physiology or Medicine 1962* (Nobel Foundation)
50	*Biographical Sketch of James Dewey Watson* (Cold Spring Harbor Lab.)
41	*A structure for Deoxyribose Nucleic Acid* (Nature, April 2, 1953)
⋮	
1	*Felix D'Herelle and the Origins of Molecular Biology* (Amazon.com)
1	Biography of Gregor Mendel
1	*Magazine: HMS Beagle Home*
1	*The Alfred Russel Wallace Page*
1	*U.S. Human Genome Project 5 Year Plan*

Stephen Hawking Community	
Score	**Site Title or Description**
85	*Professor Stephen W. Hawking's web pages*
46	*Stephen Hawking's Universe* at PBS
17	*The Stephen Hawking Pages*
15	*Stephen Hawking Builds Robotic Exoskeleton* (parody at *the Onion*)
14	*Stephen Hawking and Intel*
⋮	
1	*Did the cosmos arise from nothing?* MSNBC story
1	Spanish page for *Stephen Hawking's Universe*
1	*Relativity Group at DAMTP, Cambridge*
1	*Millennium Mathematics Project*
1	*Particle Physics Education and Information Sites*

Ronald Rivest Community	
Score	**Site Title or Description**
86	*Ronald L. Rivest : Home Page*
29	*Chaffing and Winnowing: Confidentiality without Encryption*
20	Thomas H. Cormen's home page at Dartmouth
9	*The Mathematical Guts of RSA Encryption*
8	German news story on Cryptography
⋮	
1	Phil Zimmermann's PGP web page
1	*A Very Brief History of Computer Science*
1	*Cormen / Leiserson / Rivest: Introduction to Algorithms*
1	*Security and Encryption Links*
1	*HotBot Directory: Computers & Internet, Computer Science, People: R*

Table 2: Sample results from community identification: The top five and bottom five pages (with ties) are shown for each community. The scores are the total number of inbound and outbound links that a web page has to other pages that are also in the community. Lower ranked pages often will not contain the name of the scientist used as the initial seed, yet they usually are highly topically related to the seed scientist.

Community	Most Significant Text Features
Crick	crick, nobel, dna, "francis crick", "the nobel", "of dna", watson, "james watson", francis, molecular, biology, genetics, "watson and", "structure of", "crick and"
Hawking	hawking, "stephen hawking", stephen, "hawking s", "s universe", physics, "black holes", "the universe", cambridge, cosmology, einstein, relativity, damtp, "universe the"
Rivest	rivest, "l rivest", "ronald l", ronald, cryptography, rsa, "ron rivest", lcs, "theory lcs", encryption, "lcs mit", theory, chaffing, winnowing, crypto

Table 3: The fifteen most significant text features for each community, sorted in descending order of the Kullback-Leibler metric. A feature is either a word or consecutive word-pair. To extract features, all punctuation is removed, all uppercase letters are converted to lowercase, and extra white space is removed. Although only link information is used to identify the communities, the individual pages within each community are highly topically related.

3 Experimental Results

To test the approximate community identification algorithm, we used the personal home pages of three prominent scientists as a single seed in three separate runs: Francis Crick, Stephen Hawking, and Ronald Rivest. Each trial of the approximate algorithm produced communities consisting of approximately 200 web pages. At the later stages of the runs, the induced graphs often contained tens of thousands of vertices; hence, a considerable number of web pages were pruned to produce the communities.

Table 2 shows sample web pages within the communities. On visual inspection the majority of web pages found were highly topically related and in non-trivial ways. For example, the Crick community contained many references to Darwin, the Human Genome Project, and Rosalind Franklin; the Hawking community contained many sites dealing with cosmology, relativity, and Cambridge University; and the Rivest community contained numerous encryption web sites along with sites focused on his co-authors.

Table 3 gives a more complete characterization of the three communities. We extracted all text features from the pages within a community and for ten thousand randomly chosen web pages. We then sorted all features in the community according to their ability to separate community pages from non-community pages, as measured by the Kullback-Leibler metric. Thus, the features shown in Table 3 can be interpreted as the most useful features for separating community pages from non-community pages. As can be seen, the extracted features support our hypothesis that linked-based communities are topically related.

In order to obtain more precise characterizations of the communities, we exhaustively searched for all three-term binary classifiers that disambiguate community from non-community pages. Simple disjunctive expressions of keywords related to the communities matched a large fraction of the communities with very low false alarm rates. For example, **crick** *or* **nobel** *or* **darwin** matches 54% of the Francis Crick community but only 0.5% of random web pages. Similarly, **hawking** *or* **relativity** *or* **"for mathematical"** matches 84% of the Stephen Hawking community (0.2% of random pages), and **rivest** *or* **cormen** *or* **"to encrypt"** matches 85% of the Ronald Rivest community (1.3% of random pages). The communities are strongly topically related in that they have simple and compact descriptions in the form of binary classifiers.

In comparison, simple breadth-first crawl strategies lose topical relevance very quickly. For the three scientists we investigated, only about 10% of pages at a depth of two from the seed site match the classification rules given above. In contrast, the communities that we identify have pages up to a depth of five links from the seed site. Breadth-first crawling to this depth would yield an enormous number of pages [2].

4 Conclusion

Based only on the self-organization of the link structure of the web, we are able to efficiently identify highly topically related communities, individual members of which may be spread over a very large area of the web graph. Since our method is completely divorced from text-based approaches, identified communities can be used to infer meaningful text rules and to augment text-based methods.

Applications of our method include the creation of improved search engines, content filtering, and objective analysis of the content of the web and relationships between communities represented on the web. Such analysis, taking into account issues such as the "digital divide" [9], may help improve our understanding of the world.

Acknowledgments

We thank Inktomi for the random URL data.

References

[1] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows : Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the World Wide Web. *Nature*, 401:130–131, 1999.

[3] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[4] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proc. 6th Int. Conf. on Knowledge Discovery and Data Mining*, pages 150–160, 2000.

[5] L. R. Ford Jr. and D. R. Fulkerson. Maximal flow through a network. *Canadian J. Math.*, 8:399–404, 1956.

[6] M. R. Garey and D. S. Johnson. *Computers and intractability: A guide to the theory of NP-completeness*. W. H. Freeman, New York, 1979.

[7] Andrew V. Goldberg and Robert E. Tarjan. A new approach to the maximum flow problem. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, pages 136–146, Berkeley, California, 28–30 May 1986.

[8] Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong regularities in World Wide Web surfing. *Science*, 280(5360):95–97, 1998.

[9] T.P. Novak and D.L.Hoffman. Bridging the digital divide: The impact of race on computer access and Internet use. *Science*, 281:919, 1998.

[10] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

SIDEBAR: Finding related pages on the web

Previous link-based research for identifying collections of related pages includes bibliometric methods such as co-citation and biliographic coupling [5], the PageRank algorithm [2], the HITS algorithm [7], bipartite subgraph identification [8], Spreading Activation Energy (SAE) [9], and others [6, 3].

Co-citation, bibliographic coupling, and bipartite subgraph identification are localized approaches in the sense that they seek to identify well-defined graph structures that exist inside of a narrow region of the web graph. PageRank, HITS, and SAE, are more global since they work by iteratively propagating weights through a significant portion of the web graph. The weights reflect an estimate of page importance (PageRank), how authoritative or hub-like a web page is (HITS), or how "close" a candidate page is to a starting region (SAE). PageRank and HITS are related to spectral graph partitioning [4] and therefore seek to find "eigen-web-sites" of the web graph's adjacency matrix or a simple transformation of it. Both HITS and PageRank are relatively insensitive to their choice of parameters, unlike spreading activation energy, which yields results that are extremely sensitive to the choice of parameters [9].

Localized approaches are appealing in that the identified structures unambiguously have the properties that the algorithms were designed to find. However, these approaches fail to find large related subsets of the web graph because the localized structures are simply too small. At the other extreme, PageRank and HITS operate on large subsets of the web graph and, therefore, can identify large collections of web pages that are related or valuable. However, because these methods are based on spectral graph partitioning, it is often difficult to understand and defend the inclusion of a given page in the collections that these algorithms produce. In practice, meaningful results are only achieved by HITS and PageRank when textual content is used for either preprocessing (HITS) or postprocessing (PageRank); without auxiliary text information, both PageRank and HITS have limited success in identifying collections of related pages [1].

References

[1] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proc. 21st Int. ACM SIGIR Conf.*, 1998.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. 7th Int. World Wide Web Conf.*, 1998.

[3] S. Chakrabarti, M. van der Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Proc. 8th Int. World Wide Web Conf.*, 1999.

[4] F. Chung. *Spectral Graph Theory*. CBMS Lecture Notes. Amer. Math. Soc., 1996.

[5] E. Garfield. *Citation Indexing: Its Theory and Application in Science*. Wiley, New York, 1979.

[6] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proc. 9th ACM Conf. on Hypertext and Hypermedia*, 1998.

[7] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, 1998.

[8] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. In *Proc. 8th Int. World Wide Web Conf.*, 1999.

[9] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*. ACM Press, 1996.

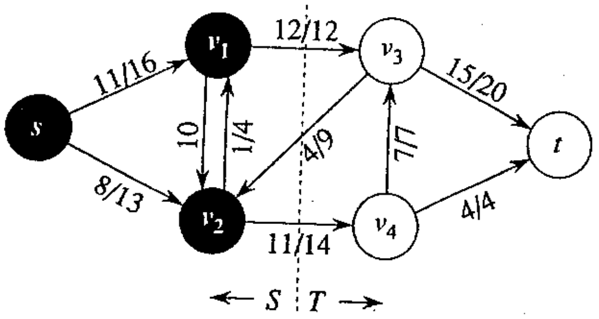


Figure 26.4 A cut (S,T) in the flow network of Figure 26.1(b), where $S = \{s, v_1, v_2\}$ and $T = \{v_3, v_4, t\}$. The vertices in S are black, and the vertices in T are white. The net flow across (S,T) is $f(S,T) = 19$, and the capacity is $c(S,T) = 26$.

we shall prove shortly, tells us that a flow is maximum if and only if its residual network contains no augmenting path. To prove this theorem, though, we must first explore the notion of a cut of a flow network.

A ***cut*** (S,T) of flow network $G = (V, E)$ is a partition of V into S and $T = V - S$ such that $s \in S$ and $t \in T$. (This definition is similar to the definition of "cut" that we used for minimum spanning trees in Chapter 23, except that here we are cutting a directed graph rather than an undirected graph, and we insist that $s \in S$ and $t \in T$.) If f is a flow, then the ***net flow*** across the cut (S,T) is defined to be $f(S,T)$. The ***capacity*** of the cut (S,T) is $c(S,T)$. A ***minimum cut*** of a network is a cut whose capacity is minimum over all cuts of the network.

Figure 26.4 shows the cut $(\{s, v_1, v_2\}, \{v_3, v_4, t\})$ in the flow network of Figure 26.1(b). The net flow across this cut is

$$\begin{aligned} f(v_1, v_3) + f(v_2, v_3) + f(v_2, v_4) &= 12 + (-4) + 11 \\ &= 19 , \end{aligned}$$

and its capacity is

$$\begin{aligned} c(v_1, v_3) + c(v_2, v_4) &= 12 + 14 \\ &= 26 . \end{aligned}$$

Observe that the net flow across a cut can include negative flows between vertices, but that the capacity of a cut is composed entirely of nonnegative values. In other words, the net flow across a cut (S,T) consists of positive flows in both directions; positive flow from S to T is added while positive flow from T to S is subtracted. On the other hand, the capacity of a cut (S,T) is computed only from edges going from S to T. Edges going from T to S are not included in the computation of $c(S,T)$.

The following lemma shows that the net flow across any cut is the same, and it equals the value of the flow.

國立中山大學九十二學年度博士班招生考試試題

科目：資訊管理論文評述第二節【資管系】

請閱讀所附論文：應用差異理論來評估學員訓練績效之研究：以商業電子化人才培訓計畫為例，並回答下列各問題。(注意：各問題有標註答題字數限制，請仔細構思你的回答在字數限制內清楚完整表達)(超過字數將扣分)

1. 所附論文共分為六節，請簡述各節的主旨 (請用中文在五百字內回答)。(10%)

2. 請就論文的內容，回答下列問題：
   (1) 研究結對於假說 H_1：受訓學員的內部培訓需求與培訓單位的外部培訓需求提供之差異間距愈小，學員感覺的培訓滿意度愈高。是被支持還是不被支持？作者根據怎樣的驗證程序來下結論。(5%)
   (2) 研究結果對於假說 H_2：受訓學員的內部培訓需求與培訓單位的外部培訓需求提供之差異間愈小，學員感覺的培訓助益愈高。是被支持還是不被支持？作者根據怎樣的驗證程序來下結論。(5%)

3. 請就論文的研究結果，回答下列問題：
   (1) 表 3 的迴歸分析結果顯示培訓提供與培訓需求對訓練效益之影響，而圖 7 與圖 8 顯示不同課程安排需求之學員在課程安排與實施滿意度之差異。如果你是這個訓練單位的負責人，會如何評估現有教學課程並做哪些改進。(請用中文在二百字內回答) (15%)
   (2) 如果訓練單位決定將所有的課程改為網路教學，要進行學員訓練績效之研究，是否仍可利用論文的研究模式與假設、變數衡量、問卷調查與資料分析技術。(請用中文在二百字內回答) (15%)

應用差異理論來評估學員訓練績效之研究：以商業電子化人才培訓計劃為例[1]

洪新原[a]　游寶達[b]　王俊程[a]　黃士銘[a]　古政元[a]　張嘉銘[a]

[a]中正大學資訊管理學系

[b]中正大學資訊工程學系

摘要

教育訓練是企業組織在維持競爭力時，必須投入資源的活動。然而，目前缺乏一套客觀的訓練評估方法，致使教育訓練難以展現其眞正價值。本研究利用 Kirkpatrick(1998) 提出訓練的四階段評估模式 (Four Levels Model)，發展出評量訓練結果的四種效益；然後結合差異理論 (Discrepancy Theory)，來解釋造成這些效益差異的原因所在。

在模式驗證的部分，我們以「商業電子化人才培訓計劃」的培訓學員爲研究對象，採用問卷調查法。調查標的則針對學員感覺的培訓滿意度、培訓助益、轉換工作意願、與轉換工作能力等效益變數、以及學員與訓練單位在課程內容、課程安排、與課程實施等方面的注重差異來加以衡量。回收的 197 份資料分析結果顯示，學員在講師的專業知識、教學態度、與表達技巧上有較高的滿意度；學員認爲所學對於公司的策略規劃上較有幫助；學員認爲受訓增加其轉換工作能力與工作機會，但想要求更好工作的意願卻沒有增加。此外，調節迴歸分析結果發現，培訓課程安排與培訓課程實施是影響學員培訓滿意的重要變數。

關鍵字：訓練評估、四階段評估模式、差異理論、商業電子化人才培訓、調節迴歸分析

[1] 本研究感謝經濟部商業司「商業電子化人才培訓計劃」的經費支持。

A Discrepancy Model and An Empirical Study of Training Evaluation

Shin-Yuan Hung
Department of Information Management
National Chung Cheng University
Pao-Ta Yu
Department of Computer Science and Information Engineering
National Chung Cheng University
Jyun-Cheng Wang, Shi-Ming Huang, Cheng-Yuan Ku, Chia-Ming Chang
Department of Information Management
National Chung Cheng University

ABSTRACT

Training is an indispensable activity for any organizations. However, a reliable method for justifying the value of training is not yet established. This study intents to develop such a method to fill the gap. We first propose four indicators to measure the training effectiveness based on Kirkpatrick's (1998) Four Levels Model. This is followed by modeling the impact of the gap between the wants of trainees and the amount they perceive is delivered by their training program. The model is based on discrepancy theory and predicts the gap is closely related to the indicators. Finally, model predictions hold true for a sample of 197 trainees of the e-Business training program. Results from moderated regression analysis indicate that both the training program design and training program implementation that trainees perceive as matching their wants have significant impact on the trainees' satisfaction.

Keywords: Training Evaluation, Four Levels Model, Discrepancy Theory, e-Business Training Program, Moderated Regression Analysis

壹、導論

許多公司組織在面臨產品或服務重新調整、員工工作類型重新調整、以及工作所需技術重新調整時，員工的在職或職前訓練成爲公司組織在維持其競爭力時，必須投入資源的活動。一般的員工訓練，會區分爲員工訓練與員工發展，兩者在效益實現的時間雖然有所不同，也就是前者比較重視立即可用，而後者比較重視未來的工作發展。但是，所希望得到的結果都是一樣的，亦即透過訓練來增進員工的能力，使其在工作上可以更具效能。此外，在面臨經濟不景氣或經濟轉型時，許多待業與失業的人口也需要進階或第二專長訓練，來幫助他們找到合適工作與發展事業第二春。儘管訓練 (Training) 對於組織與個人是如此的需要，然而目前卻缺乏一套評估訓練績效的方法，可以用來說明這些投資的效益。

Kirkpatrick(1998) 於 1959 年提出了四階段評估模型 (Four Levels Model)，他認爲訓練的績效評估應分爲四階段來衡量：第一階段爲滿意度評估；第二階段爲爲測試評估；第三階段爲工作改善評估；第四階段爲組織改善評估。這個四階段評估模型，可以應用來發展評量訓練成果的效益。另外，差異理論 (Discrepancy Theory) 指出，對於不同效益的滿意差異，是來自受測者本身比較其實際經歷與期望標準之間的差異所產生 (Locke, 1976; 1969)。差異理論的應用，在過去已有諸多例子。例如：資訊部門員工的工作滿意與離職意願 (Jiang and Klein, 2001)、工作生活品質的滿意度 (Wilcock and Wright, 1991)、大學畢業生在畢業前後對於工作期望與滿意度的比較 (Fricko and Beehr, 1992)、工作滿意度的探討 (Rice et al., 1989)、供需之間對產品品質的認知差異 (Jacques and Taylor, 1995)、以及銀行職員與客戶對於服務的認知差異 (Gilles, 1995)。這個理論可以應用來解釋不同效益的差異原因。

因此，本研究利用 Kirkpatrick(1998) 提出的四階段評估模式，發展出用來評量訓練成果的培訓滿意、培訓助益、轉換工作意願、及轉換工作能力等四種效益；然後結合差異理論 (Locke, 1976; 1969)，藉由比較其實際經歷與期望標準之間的差異，來解釋造成這些效益差異的原因所在。至於用來實證這套訓練評估方法的資料，是以我們在去年所承辦的「商業電子化人才培訓計劃」這個全國性大型商業電子化人才培訓課程爲研究對象，透過對受訓後學員的問卷調查，蒐集到 197 份有效問卷來從事分析，以實際驗證這個方法的有效性。

經由本研究所提出訓練評估方法的實行，我們可以評估受訓學員對於培訓課程的滿意度、評估培訓課程對於學員公司的助益、測試培訓課程對於學員轉換工作的意願、以及瞭解培訓課程對於學員工作轉換能力的增進。此外，我們也可以瞭解影響這些不同培訓課程績效的因素，以提供未來培訓課程設計與編修的依據。

本文的結構，依序爲第二章探討訓練的設計與實施、訓練評估的相關研究、以及差異理論的相關應用等理論與文獻；第三章描述研究對象，說明「商業電子化人才培訓計劃」的規劃、進行、與目前成果；第四章建立研究模式、形成研究假設、並且說明研究方法；第五章結果分析與討論；最後則提出結論與建議。

貳、文獻探討

一、訓練的設計與實施

所謂訓練 (Training) 是指組織爲適應

業務及培育人才需要，對所屬員工予以有計畫的增進所需學識技能，減少個別差異，以期員工能勝任現職工作，及將來擔任更重要職務。訓練種類的區分，傅肅良(1985)指出員工訓練的種類，可區分爲職前訓練、在職訓練、職外訓練等三種。首先，職前訓練是指組織對遴選新進人員在派職前所舉辦之訓練，可區分爲一般性或專業性的職前訓練。其次，在職訓練是指組織員工於在職期間，參加由組織或其上級組織所舉辦之訓練，在訓練期間多爲帶職帶薪。在職訓練依其性質與目的之不同，又可區分補充學能訓練、儲備學能訓練、人際關係訓練、與運用智慧思考訓練。最後，職外訓練則是組織的員工，暫時離開現職及處所，至相關學術機構參加長期的訓練。參加此種訓練者，視其受訓期間長短，可爲留職停薪或帶職帶薪。職外訓練依其性質又可分在校進修，在機構實習、以及赴各機構考察等。

在設計訓練時應注意下列五點：(1)不同的訓練內容應適用不同的學習理論：如以學習技術及操作爲主的訓練，應用刺激反應的聯結論，較能說明學習的意義與原則；對以學習思考及解決問題爲主的訓練，則以應用符號完形的認知論，較能說明學習的意義與原則。如根據訓練內容所選用的學習理論，來設計訓練的計畫，將可增進訓練效果；(2)不同的訓練教材應運用不同的學習方法：如以技術及操作爲主要內容的教材，於施訓時，宜採用實地操作方法學習，並略增加學習的次；數，以加強刺激與反應間的聯結，進而增強學習的效果。對以思考及解決問題爲主要內容的教材，宜採用討論個案研究等方法施訓，並以學識較具水準的員工參加，以增進學習效果；(3)參加訓練員工事先需有心理準備：所謂心理準備是指參加訓練的員工對於訓練計畫、方式、及方法，應有相當瞭解；對舉辦訓練的目的及訓練後的期望，應有所認識。員工有了此種瞭解與認識後，在心理上比較能接受訓練的教材與接受指導人員的指導；(4)對學習具有反應的員工應即予適當的獎勵：員工參加訓練，如對訓練方面表現有良好成績者，應該給予適當的獎勵，員工一旦獲得獎勵，更會增加其學習興趣與動機，會因而表現出更好的學習績效，如此將會增進學習效果；(5)使員工迅速獲知學習的成績，俾有助於改正缺失：員工在訓練期間的學習成績，必須使員工能迅速的獲知，惟有如此，員工始能及時改正自己學習上的缺點，及加強與發揚自己學習上的優點。

傳統上的訓練環境可以時間、地點、空間等三個構面來定義。其中，時間是指訓練的時間；地點是指訓練的發生地點；空間則是指教材與學習資源的獲取。Piccoli et al. (2001) 進一步指出，虛擬訓練環境與傳統訓練的差異，還額外包含：協助學員參與的教材呈現工具（技術）、學員間與師生間的溝通（互動）、以及學員控制課程內容的呈現（控制）。

本研究對象爲參加「商業電子化人才培訓計劃」的學員，學員的參與屬於職外訓練，在整個訓練規劃程序中，對學員學習影響的主要因素，包含訓練教師的教學效率、學員的學習成就、訓練課程的設計、訓練時間、訓練地點、以及準備用具與資料等，並以學習的影響因素與教學評量類別區分，參酌訓練課程設計專家的意見，修改適合本研究對象之影響學習因素類別分類及衡量問項，即商業電子化人才培訓學員訓練績效的三大影響因素爲：(1)培訓課程內容：包含課程內容結構的安排、教材與教案的印刷與編排、以及教材與教案的提供方式；(2)培訓課程安排：包含學習測驗的舉行方式、學習測驗的難易度、訓練場地與設施的安排、以及上課時間的分配與安排；(3)培訓課程實施：包含講師的表達技巧、講師的專業知識、講師

的教學態度、以及課程對其工作上的幫助程度。

二、訓練評估的相關研究

(一)訓練評估的需要性

舉辦訓練，花費人力、時間與經費甚多，故必需使訓練有效果，否則是極大的浪費。訓練之有無效果，通常透過評估方法來衡量。然而在實務界，許多組織並未收集資料，以決定他們自己的計畫成效 (Goldstein, 1992) 。 Ralphs and Stephan (1986) 曾針對五百大明星企業做過調查，發現大部份企業（佔 86% 強）的評鑑方法均是受訓者在課程結束後填寫回應，很少有企業在受訓後收集受訓者在工作表現上的改變。 Sarri et al.(1988) 調查超過六百家的企業，獲得相似的結果。而在國內，余品嫻 (1997) 指出國內目前許多訓練，評估的缺乏應可說是最普遍的缺失。訓練人員經常假定訓練方案有其價值而避免從事評估，或只著重受訓者對訓練之感受與經驗此種反應性評估，而對訓練內容本身並不加以檢視，原因在予「訓練評估」的高困難度、冗長乏味及費時極長、有許多變項可能干擾評估的準確度、評估所得的結果也容易招致外在的批評或壓力、有時評估的結果也不一定使人信服等。這些困難均使得訓練人員傾向於訓練可以運作即可，而儘可能避免去從事此種吃力不討好的工作。

(二)訓練評估的時機與方法

訓練評估的時機有兩種（傅肅良，1985）：(1)訓練結束時評估：於訓練結束時，對參加訓練人員在訓練期間的各種表現作成評估，並與未參加訓練前的表現相比較，以認定舉辦訓練有無成效。訓練結束時評估的重點，包括：有受訓員工的學識有無經由訓練而增進及增進多少，受訓員工技能有無獲得及獲得多少，受訓員工工作態度有無改善及改善多少，受訓員工工作情緒有無提高及提高多少，受訓員工參加會議或工作討論會時發言是否踴躍及是否有價值，員工在受訓期間是否已充分的接受訓練等；(2)訓練結束回任工作後評估：訓練的目的，不在員工於受訓期間有無表現，而在結訓回任工作後在工作上有無表現。故結訓回任工作後的評估，要比訓練結束時的評估還要重要。回任工作後的評估重點，主要包括：工作態度有無改變及改變之程度如何與維持的期間有多久，工作效率有無增進及增進之程度如何，對訓練的目標有無達成等。

至於訓練評估的方法，傅肅良 (1985) 指出，訓練結業時評估可採用下列方法：(1)應用學識技能的測驗評估訓練成效；(2)應用工作態度調查評估訓練成效；(3)調查員工有關訓練的改進建議；(4)記錄訓練期間出席人員變動情形；(5)根據觀察員所提報告評估訓練成效；(6)根據主持、指導及協助訓練人員的報告評估訓練成效；(7)從受訓員工的結訓成績評估訓練成效。訓練結束回任工作後評估則可採取下列方法：(1)結訓後每隔相當期間，以調查受訓員工的工作效益評估訓練成效；(2)調查或訪問受訓員工的上級主管或下屬，根據所得意見評估訓練成效；(3)實地觀察結訓員工的工作實況評估訓練成效；(4)分析結訓員工的人事記錄評估訓練成效；(5)根據曾受訓與未受訓員工工作效率比較評估訓練成效；(6)根據曾受訓員工能否達到工作標準評估訓練成效；(7)根據訓練目標的有無達成評估訓練成效。

(三)訓練評估的指標

Kirkpatrick(1998) 對於訓練成效的評估，提出四階段評估模型，廣爲業界所採用。所謂四階段主要指：(1)反應階段：即受訓者對訓練的反應，此種反應資料可作

爲決定下次訓練計畫的參考。此種反應資料多於訓練結束後，以問卷或面談的方式取得。它包括受訓者是否喜歡訓練計畫？喜歡的程度如何？(2)學習階段：此即爲受訓者對所授內容、原則、觀念、知識、技能與態度等的學習程度。該項標準在於獲得受訓者對課程的學習方面，而不在於是否能運用所學於工作上。對於知識、內容、原則、觀念等，可以各項測驗或考試方法，加以評估。對於技能可以實作測驗評估，至於態度則可施予態度量表評估之。(3)行爲階段：此爲評估受訓者在接受訓練後，工作行爲改變的程度。有關用手操作的工作，運用有系統的觀察法，即可得到完整資料。但對複雜的工作，就得分別採用其他適當方法，如工作抽查、主管考核、自我評核、自我記錄等，才能得到完整的工作行爲資料。當然，上述各種方法也可綜合運用。(4)結果階段：此乃爲評估受訓者的工作行爲，是否影響到組織功能的實施，最後的結果是否已達成？這些評估範圍包括工作效率、成本費用、生產質量、員工流動、態度改變、目標認知、業務改進等。結果標準評估的困難，乃爲如何認定這些改變是訓練成果所形成的。蓋效能的提高和經驗也有關係，工作的進步是經驗與訓練的共同作用。如果經過訓練後，工作成效有立即的改變，才能確定訓練的價值，否則很難正確地評估出訓練的成效。

林欽榮 (1997) 指出評估指標的選擇，應按訓練目標而定，該四項指標可以共同評估，但也可個別評估。理想上，最好對此四項指標都各自設定一個目標，予以評估。總之，目標訂得愈精確，訓練評估也愈正確。因此，訓練評估指標，即爲目標設定的標準。

在訓練評估的時機與方法上，本研究採用訓練後的評估，樣本中的這些受訓後的學員，絕大多數距離受訓完都超過三個月以上的時間，有充足的時間讓他們消化吸收，並讓訓練的效益實現出來。至於訓練評估的指標，則是參考 Kirkpatrick (1998) 的架構，以對應的培訓滿意、培訓助益、轉換工作意願、及轉換工作能力來衡量其效益。

三、差異理論的相關應用

差異理論 (Discrepancy Theory) 應用在訓練評估的解釋時，是指訓練效益的不同（在差異理論中稱爲 Facet ）決定於學員心理對於實際培訓提供（稱爲 Facet Amount ）與預期受訓需求（稱爲 Wanted Amount ）的比較所產生的差異（稱爲 Perceived Have-Want Discrepancy ），這個心理比較的過程可能產生正的差異或負的差異。當學員感受到培訓提供大於他的預期需求時，會有正的差異產生；反之，則會有負的差異產生。不論這種差異爲正或爲負，只要其夠大時，就會顯著地影響受訓學員感受到的訓練效益。至於正負差異影響訓練效益的效果，則視此訓練效益的內容而定。舉例來說，學員受訓後，倘若覺得培訓單位所設計的培訓課程內容不如預期，就會產生負的差異，這個負的差異過大時，就會造成學員對於培訓課程的感覺不滿意；反之，如果培訓單位所設計的培訓課程內容符合學員的預期，那麼學員就會有比較高的滿意度。

過去差異理論應用在個人差異影響的探討相當多，例如：Jiang and Klein(2001) 以差異理論模型，建立資管部門員工工作滿意與離職意圖的預測模型。研究結果顯示，在員工對工作期望－實際水準的差距值愈大時，員工工作滿意程度愈低與離職意願愈高等。Cooper and Artz(1995) 以差異理論爲理論架構，探討企業家個人滿意。滿意的決定由實際績效報酬與個人預期目標之間的差異來決定。從 287 位企業家，進行 3 年的縱向研究，研究結果發現

企業家強調的非經濟性目標比起經濟性目標，更令企業家滿意，而企業家初期的期望也與滿意水準成正相關。此研究結果可以說明爲企業家在面對類似的財務績效時，對其投資會有不同程度的滿意。

此外，Johnson and Petrie (1995) 以差異理論來探討性別差異的眞實與理想知覺，對飲食異常行爲和態度的可能影響。研究結果顯示對厭食與食慾過盛沒有性別角色差異的女性，較不關心體型，但有較高度自尊。Thompson and Bunderson (2001) 指出過去大多數的研究對工作與非工作的時間衝突都著重在時間的分配，強調時間分配的平衡。平衡的意像受限於其忽略時間的知覺經驗與人支配上的主觀意義。本研究從角色認同與自我差異理論探討，發現工作與非工作時間的衝突不僅是因時間量的關係，也是因爲工作與非工作時間之間認同差異的關係。

經由上述對於差異理論的探討，我們可以瞭解造成學員對訓練效益（包括：培訓滿意、培訓助益、轉換工作意願、及轉換工作能力）感覺不同的原因，可以從學員對於各個影響因素（包括：培訓課程內容、培訓課程安排、及培訓課程實施）在實際培訓提供與預期培訓需求的差異程度，來提出合理的解釋。

參、研究對象描述

本研究的對象選定我們在 90 年度所承辦「商業電子化人才培訓計劃」的受訓學員。該計劃乃是政府有鑑於目前國內企業電子化人才嚴重不足，而且由於經濟不景氣與面臨轉型，許多的待業人才也需要第二專長的輔導，以利其轉業與就業。爲了因應這股強烈的需求，經濟部商業司特別在去年開闢該計畫，以培育企業電子化管理人才，同時建立國內企業電子化教育環境的基礎。

該計劃共開設三類培訓課程，包含「商業電子化營運作業管理」、「商業電子化專案管理」及「商業電子化策略規劃」之培訓課程。每項培訓課程均訂出四項策略來逐步落實，包含培育課程之推展、核心教材之撰寫、網頁教材之建置、績效之評估與分析（該計畫之目標規畫如圖 1 所示）。首先，在培育課程之推展上，在全國北中南各地同步開設培訓課程，成效卓著，各地皆反應熱烈，本年度共計培訓 795 人次。其次，在核心教材之撰寫上，編輯流程融入品保觀念與流程控管，由國內學者專家及相關業界人士組成審議小組，審查所編訂之教材及課程規劃。對於教材內容，每本書皆透過書面審查，並召開審查會議，依據審查意見修改教材格式及內容，並提供不同形式之教材，包含出版「商業電子化營運作業管理」、「商業電子化專案管理」及「商業電子化策略規劃」之網路教材、書籍、及光碟版本共三套。再者，在網頁教材之建置上，完成建構電子商務的網路學習網站(網址: http://www.ebusiness.ccu.edu.tw)，提供民衆透過網路下載所開設的電子商務訓練課程的相關資料。最後，在訓練績效之評估與分析上，爲有效落實所有培訓課程的績效評估，我們發展一套訓練評估之制度與方法，以作爲培訓計畫中學員學習績效之評估。

商業電子化系列課程總計開設電子化營運作業管理班、電子化專案管理班、以及電子化策略規劃班等三個班別，在北中南三區共開立 18 個班次。其中，電子化營運作業管理班總計培訓 422 位學員；電子化專案管理班總計培訓 262 位學員；電子化策略規劃班總計培訓 110 位學員。根據這些學員的基本資料分析，我們得到以下結果：(1)依性別統計：全部學員人數總計爲 794 人，女性人數有 290 名，占 37 %，男生人數爲 504 名，占 63 %。參加

受訓的學員中，男學員的比例較高；(2)依年齡統計：全部學員的年齡分佈爲 16 至 20 歲有 7 人、21 至 25 歲有 98 人、26 至 30 歲有 212 人、31 至 35 歲有 177 人、36 至 40 歲有 124 人、41 至 4 5 歲有 91 人、46 至 50 歲有 52 人、51 歲以上的有 33 人。受訓學員的年齡層次集中在 21-40 歲，也就是青壯年的學員居多；(3)依行業統計：受訓學員工作的行業別依序爲服務業（有 281 人，占 35 ％）、製造業 (143 人，占 18 ％）、資訊業 (86 人，占 11 ％）、教育業 (82 人，占 10 ％）、政府機關 (45 人，占 6 ％）、以及金融業 (29 人，占 4 ％）等。另外，還有 35% 的受訓學員是屬於待業中。

肆、研究方法

一、研究模式與假設

根據文獻探討的結果，培訓課程績效評估的相關因素可分爲下面二類，即自變數：(1)學員的內部培訓需求、(2)訓練課程的外部培訓提供。因變數則包含：(1)培訓滿意、(2)培訓助益、(3)轉換工作意願、(4)轉換工作能力。本研究的目的除了瞭解受訓學員對於培訓滿意、培訓助益、轉換工作意願、轉換工作能力等訓練效益的感覺外，我們也進一步要探討學員對於課程內容、課程安排、課程實施等的需求與感受

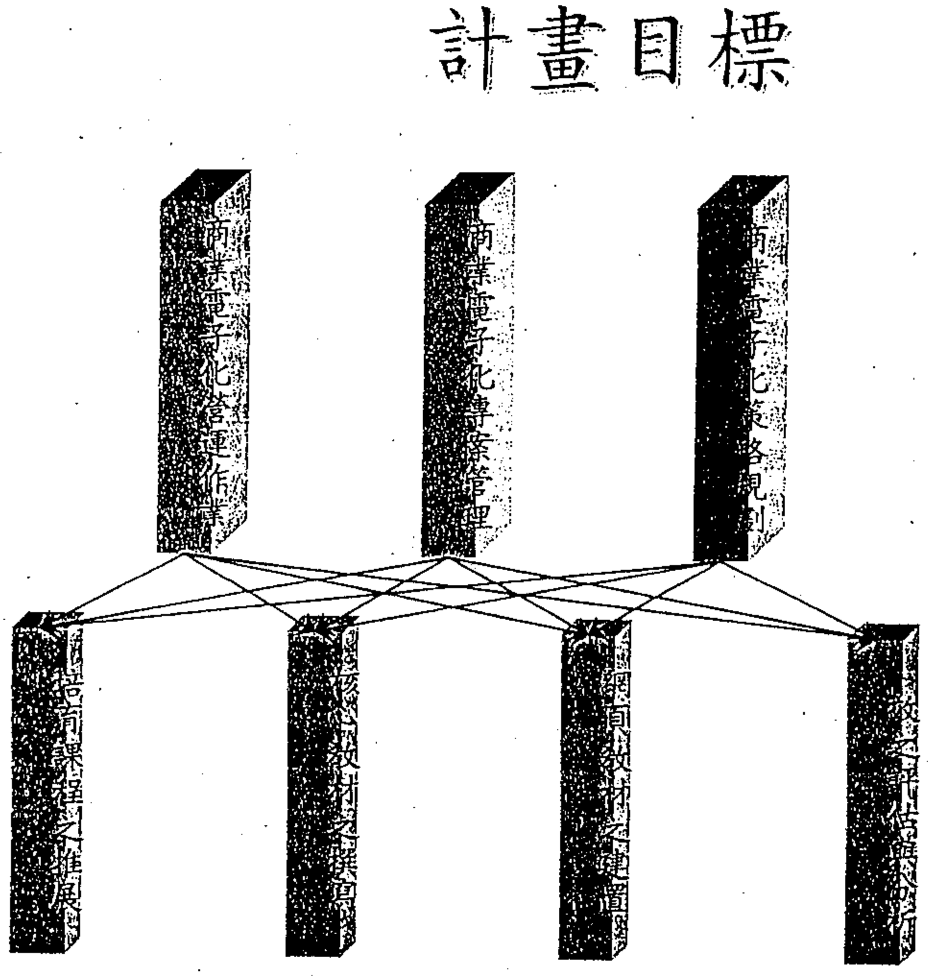


圖1：商業電子化人才培訓計劃目標及規劃

到訓練單位的提供程度之間差異間距，對於訓練效益感覺的影響。因此，我們參考差異理論與相關實證研究結果，建立的研究模式（見圖 2）與研究假設如下所示。

H1: 受訓學員的內部培訓需求與培訓單位的外部培訓提供之差異間距愈小，學員感覺的培訓滿意度愈高。

H2: 受訓學員的內部培訓需求與培訓單位的外部培訓提供之差異間距愈小，學員感覺的培訓助益愈高。

H3: 受訓學員的內部培訓需求與培訓單位的外部培訓提供之差異間距愈小，學員愈能提高其自信心，進而其轉換工作意願可能愈高。

H4: 受訓學員的內部培訓需求與培訓單位的外部培訓提供之差異間距愈小，學員愈能感覺到轉換工作能力的提高。

二、變數衡量

本研究量表包含內部培訓需求、外部培訓提供、培訓滿意、培訓助益、轉換工作意願、轉換工作能力四項構念（問卷內容請參考附件一）。量表設計的主要參考來源限制在教育訓練員工爲實証對象之研究所使用過的問卷，再以主要量表的參考量表爲輔，最後我們與教育訓練之課程編製專家逐條審閱量表各構念題項而定稿；整體而言，量表取材自與本研究對象相似的問卷，並參酌學界教育課程教授與業界人力資源訓練主管的意見修訂而成，應具

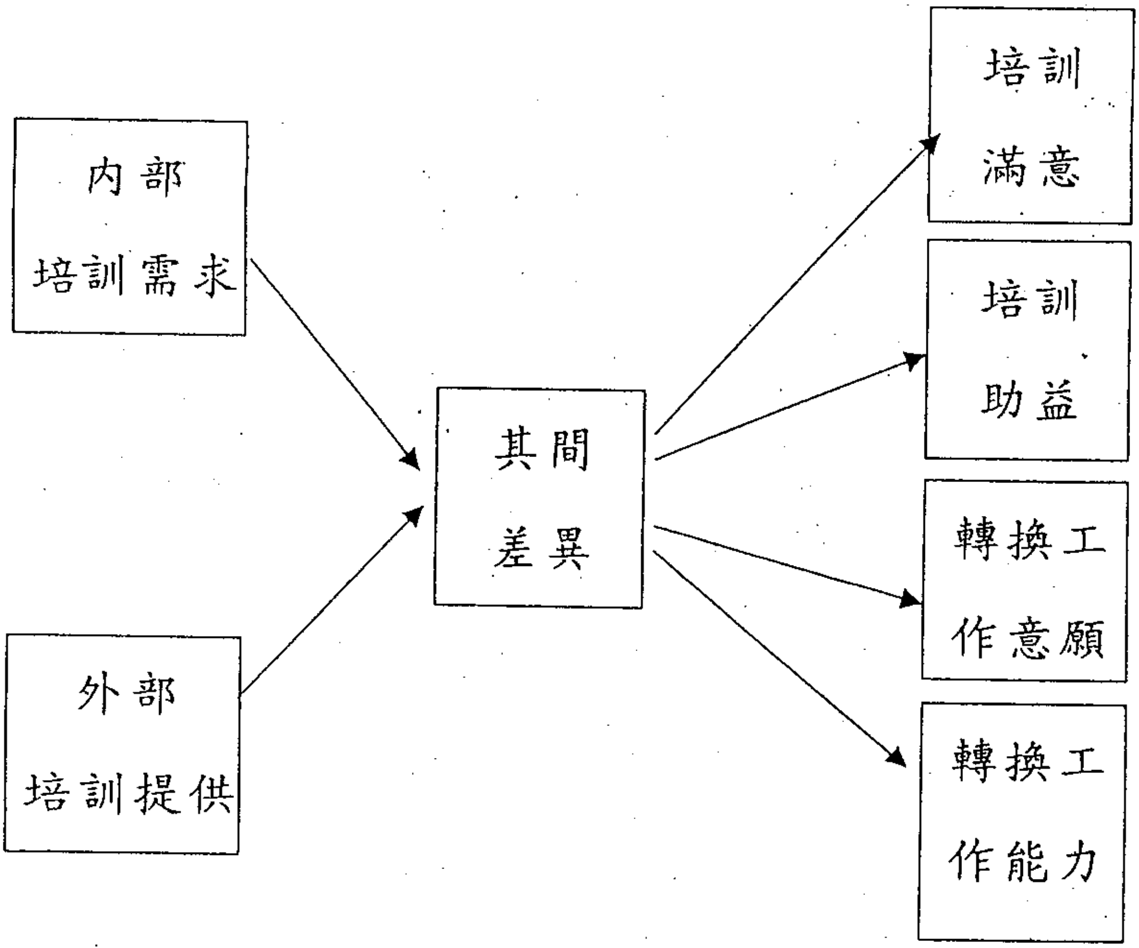


圖2：研究模式

備內容效度，並可適用於本研究中的受訓學員。至於，各個構念指標項目的衡量則皆以5點Likert尺度來量測其幅度大小，從1代表「非常不同意」到5代表「非常同意」。以下分別說明各構念量表來源：

㈠內部培訓需求

我們採用三個構面：課程內容（課程內容結構的安排、教材與教案的印刷與編排、教材與教案的提供方式）、課程安排（學習測驗的舉行方式、學習測驗的難易度、訓練場地與設施的安排、上課時間的分配與安排）、課程實施（講師的表達技巧、講師的專業知識、講師的教學態度、課程對我工作上的幫助程度），共十一個量測題項來衡量學員內心對培訓課程的需求。

㈡外部培訓提供

同樣地，衡量學員對培訓單位感受到的提供程度，我們亦採用相同三個構面：課程內容（ 課程內容結構的安排、教材與教案的印刷與編排、教材與教案的提供方式）、課程安排（ 學習測驗的舉行方式、學習測驗的難易度、訓練場地與設施的安排、上課時間的分配與安排）、課程實施（講師的表達技巧、講師的專業知識、講師的教學態度、課程對我工作上的幫助程度），共十一個量測題項來衡量學員所感受到培訓單位對於這些構面的提供程度。

㈢培訓滿意

我們直接採用參考一般學界與業界對於培訓課程的評估重點與項目所發展出來，而在實證研究上，此量表已被廣泛使用過。本研究共採用十一個量測題項來衡量培訓滿意。

㈣培訓助益

培訓助益題項係依照企業組織的營運，管理，與規劃不同層面來評估培訓課程對於學員在處理公司事務上的助益。本研究共採用三個量測題項來衡量培訓助益。

㈤轉換工作意願

轉換工作意願題項係採用Khatri et al.(2001)的量表，這份量表共包含三個量測題項來衡量離職意願(Turnover Intention)的構面。本研究採用這三個量測題項來衡量轉換工作意願。

㈥轉換工作能力

轉換工作能力量表，我們採用Harris and Fink(1987)的量表，這份量表共包含三個量測題項來衡量感受到的其他工作機會(Perceived Alternatives)的構面。本研究採用這三個量測題項來衡量轉換工作能力。

三、調查流程

為瞭解商業電子化系列課程各位學員的學習效果，本研究先以若干名受過培訓之學員為預試對象，以了解本研究的問卷內容是否為受測者所瞭解，經修改後，正式寄出問卷。正式研究以參與商業電子化人才培訓計劃課程之學員為對象，利用郵件寄發問卷內容，資料蒐集時間為2001年10月19日至11月21日。在扣除找不到人與郵寄退回者後，總計寄發出去的問卷共有498份，扣除回答不完整之無效問卷，共計回收有效問卷有197份，回收率為40％。

四、資料分析技術

本研究所使用的資料分析技術，包括：敘述統計分析與調節迴歸分析(Moderated Regression Analysis)。敘述統計的平均數與變異數，用來讓我們瞭解受訓學員對於培訓滿意、培訓助益、轉換

工作意願、以及轉換工作能力等訓練效益的感覺。調節迴歸分析則是應用來檢定假設，瞭解「提供-需求」差異間距 (Gap) 對於四個訓練效益變數的預測能力。調節迴歸分析的分析流程有三個步驟：首先，提供及需求當成二個自變數，而訓練效益當成因變數來得到第一條迴歸方程式；其次，提供、需求、以及提供與需求的交互作用項當成三個自變數，而訓練效益當成因變數來得到第二條迴歸方程式；最後，比較第二條與第一條迴歸方程式解釋變異能力（ R^2 ）的增加顯著與否，如果有顯著增加的話，表示提供與需求的間距對於訓練效益變數有顯著的預測能力；反之則無。

伍、結果分析與討論

一、回收樣本特性與代表性

本研究的回收樣本特性如表 1 所示。樣本的代表性可由以下三個角度來審視：首先，我們比較樣本與母體的性別、年齡、職業等基本資料，發現這些項目的分布相似（見表 1）；其次，經由對訓練效益變數的分析，我們發現樣本中的受訓學員，未見過度集中於二尾的極度正面或極度負面效益上；最後，經由對表 1 中所列的基本資料變數進行變異數分析後，結果發現這些基本資料變數對於訓練效益變數的影響都不顯著。以上這些分析結果，可以說明本研究所得到的樣本，具有足夠的母體代表性。

二、訓練效益

受訓學員所感受到的訓練效益如圖 3 至圖 6 所示。我們從這些結果可以發現：

1. 在培訓滿意度上（見圖 3），整體來說，受訓學員對於培訓課程還滿意。其中尤以講師的專業知識、教學態度、與表達技巧等方面，學員滿意度最高。我們所聘請的講師陣容（可參考計劃網站，網址：http://www.ebusiness.ccu.edu.tw）包含各大學中學有專精的教授，以及業界中代表性公司的高階主管，因此深獲好評。而教材與教案的印編與提供方式、學習測驗的舉行方式與難易度等方面則有改善的空間。我們提供教材與教案的方式，是讓學員經由網站下載，所以可能由於頻寬的問題，導致速度過慢或無法下載。此外，由於完整的教材印編程序冗長（包括：初稿、一校、二校、三校等步驟），因此基於上課的需要，而以未經校稿的初稿當作上課教材，內有許多筆誤與編排不全處，可能造成學員印象不好之處。
2. 在培訓助益上（見圖 4），整體來說，受訓學員認為參加這個訓練課程後，對於他們將來可為公司貢獻更多，尤其是在策略規劃方面。學員感覺受訓所學對於公司的幫助程度，最高的是高階的規劃工作、其次是中階的管理工作、最低的是低階的營運作業。雖然受訪對象中，反而是策略規劃班的學員數最少、其次是專案管理班的學員數、最多的是營運作業班的學員，然而我們在營運作業班的課程設計上，除了有一半技術的課，另外也有一半是管理的課，可以讓這些學員瞭解到電子商務對於公司策略規劃上的幫助。
3. 在工作機會與轉換能力上（見圖 5），受訓學員感覺參加這個訓練課程後，可以增加他們的工作機會，同時也增加他們轉換工作的能力。至於，想要求更好工作的平均分數不

表1：回收樣本特性與母體之比較

	樣	本	母	體
	人 數	百分比	人 數	百分比
性別				
男	130	66.00%	504	63.48%
女	66	33.50%	290	36.52%
年齡				
20歲以下	1	0.50%	7	0.88%
21-30歲	77	39.10%	310	39.04%
31-40歲	84	42.60%	301	37.91%
41-50歲	31	15.70%	143	18.01%
51歲以上	4	2.00%	33	4.16%
學歷				
高中職	9	4.60%		
大專	141	71.60%		
研究所或以上	45	22.80%		
職位				
一般職員	106	53.80%		
小主管	38	19.30%		
中階主管	33	16.70%		
高階主管	6	3.00%		
婚姻				
已婚	86	43.70%		
未婚	106	53.80%		
職業				
金融業	14	7.10%	29	3.65%
軍公教	45	22.80%	127	15.99%
資訊業	40	20.30%	86	10.83%
服務業	28	14.20%	38	4.79%
醫療業	4	2.00%	15	1.89%
運輸/旅遊	1	0.50%	15	1.89%
娛樂/出版/傳播/行銷	10	5.30%	25	3.15%
農漁牧	2	1.10%	2	0.25%
製造業	21	10.65%	143	18.01%

高的原因，可能是由於經濟不景氣，失業率攀升，因此學員會比較保守，而著重工作的安全與穩定。

4. 在學員與訓練單位注重層面的差異上（見圖 6），學員和訓練單位在講師的表達技巧、教學態度、以及教材教案的印刷編排上的注重程度有明顯差異。而在學習測驗的難易程度與舉行方式上，學員與訓練單位的重視程度則無太大差異。由表 2 中的統計數字，我們可以瞭解學員和訓練單位在講師的表達技巧與教學態度都相當注重（分列所有 11 項中的第 2 和 3 名）。因此，雖然學員的需求仍高於訓練單位的提供，學員在這些項目的滿意度還是相當高。至於教材教案的印刷編排，在學員的感覺是所有 11 項裏最低的，因此滿意度也排名在後。

三、假設檢定

爲了檢定研究假設，我們針對每個訓練效益變數，進行三次的調節迴歸分析，

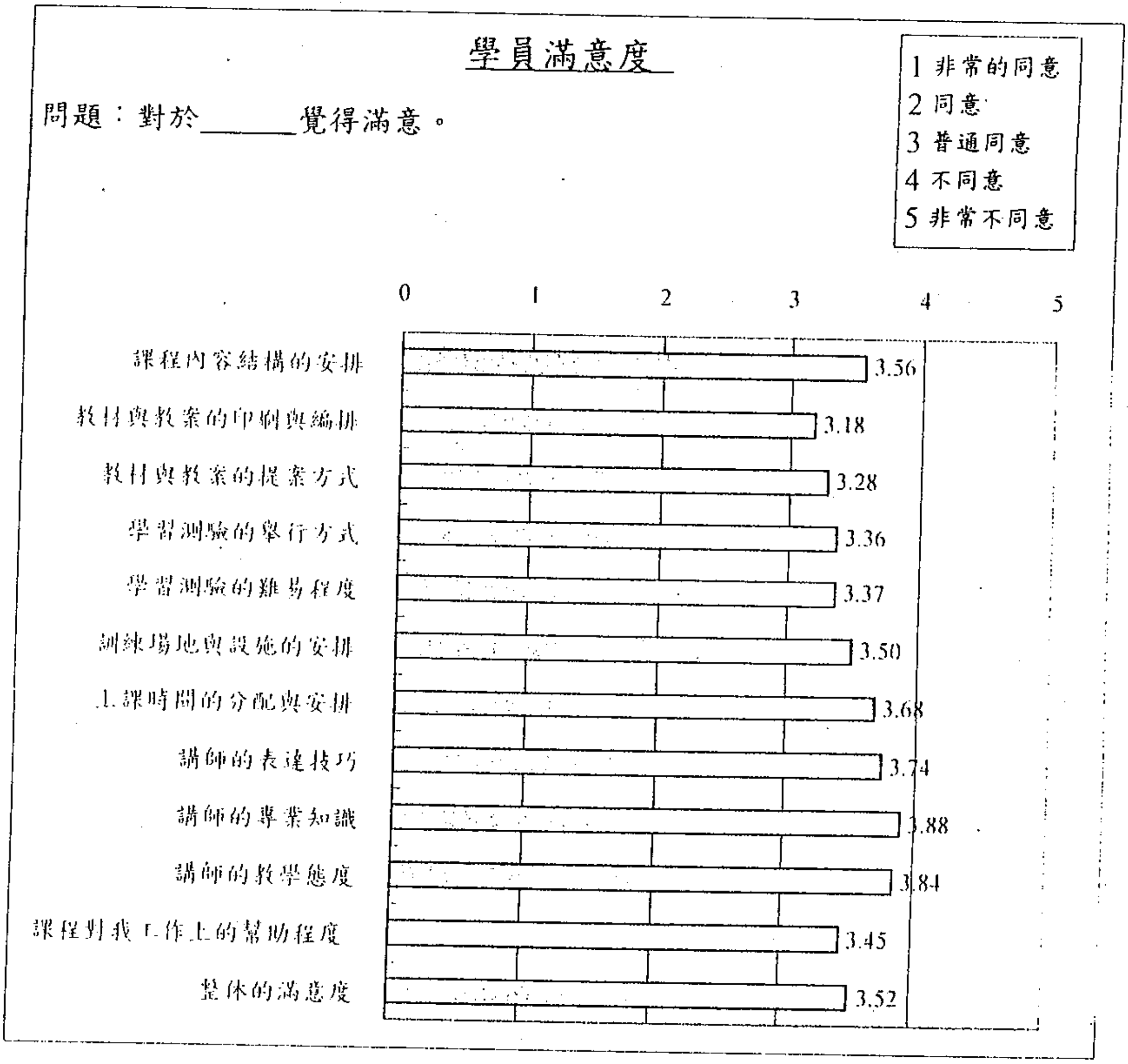


圖3：培訓滿意的調查結果

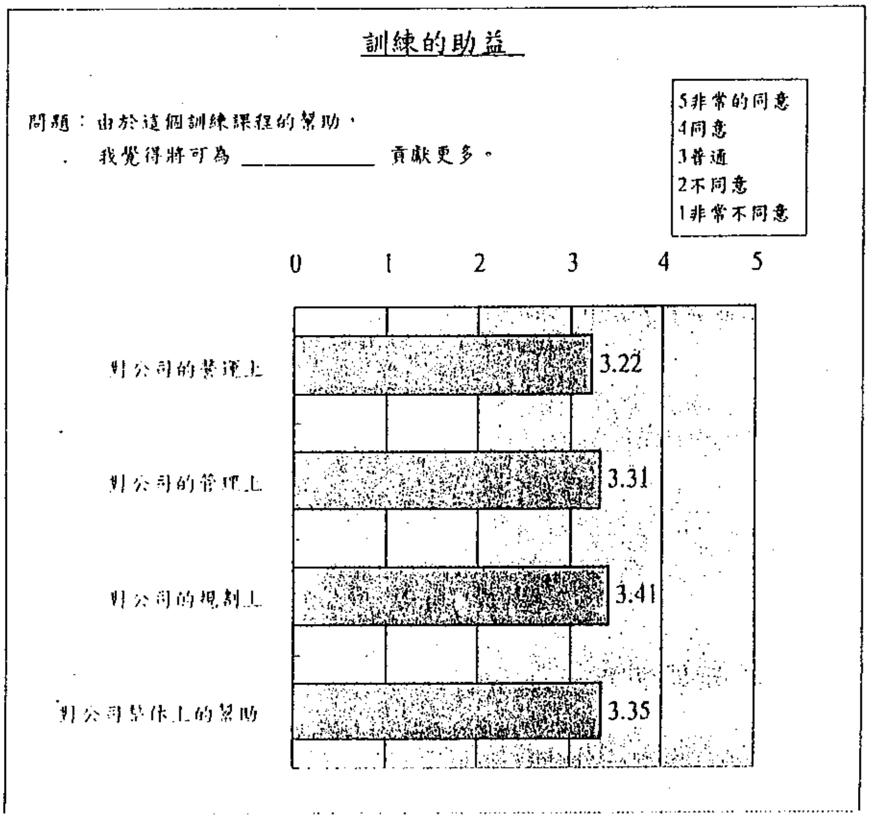


圖4：培訓助益的調查結果

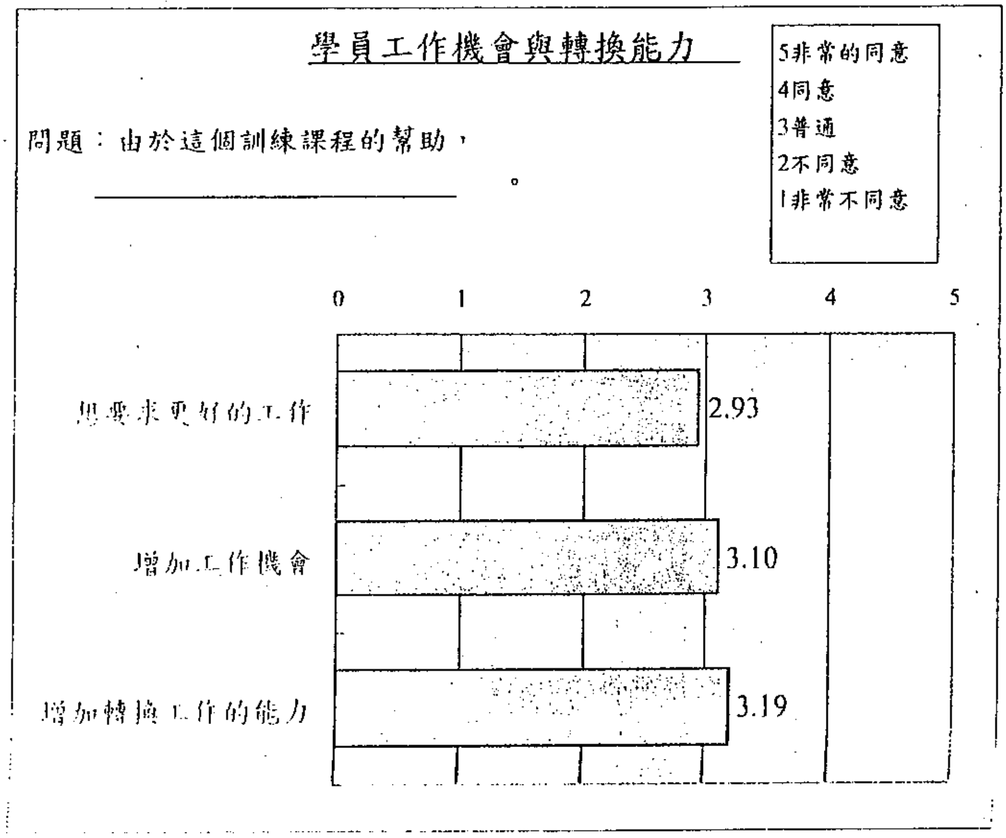
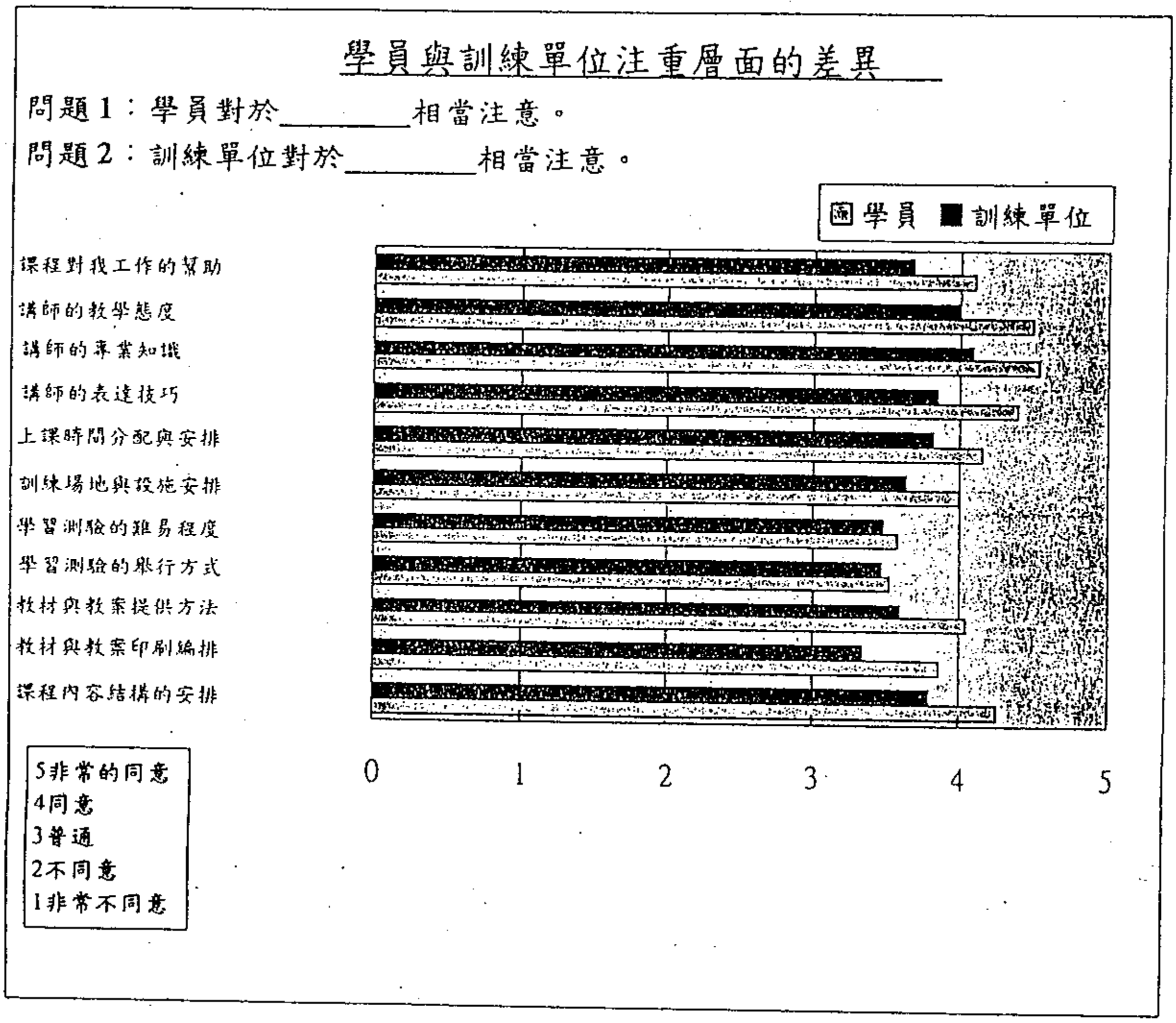


圖5：工作機會與轉換工作能力的調查結果



學員與訓練單位注重層面的差異

問題1：學員對於________相當注意。

問題2：訓練單位對於________相當注意。

圖6：學員與訓練單位注重層面差異的調查結果

表2：培訓需求與培訓提供之差距

構面		外部 培訓提供	內部 培訓需求	差距 (提供-需求)
培訓課程內容		3.57	4.05	-0.47
	課程內容結構的安排	3.79(5)	4.24(4)	-0.46
	教材與教案的印刷編排	3.34(11)	3.85(9)	-0.52
	教材與教案的提供方法	3.59(8)	4.04(7)	-0.45
培訓課程安排		3.59	3.81	-0.22
	學習測驗的舉行方式	3.46(10)	3.52(10)	-0.06
	學習測驗的難易程度	3.47(9)	3.58(11)	-0.11
	訓練場地與設施的安排	3.62(7)	3.99(8)	-0.37
	上課時間的分配與安排	3.81(4)	4.15(5)	-0.34
培訓課程實施		3.89	4.37	-0.48
	講師的表達技巧	3.84(3)	4.39(3)	-0.55
	講師的專業知識	4.08(1)	4.53(1)	-0.45
	講師的教學態度	3.98(2)	4.48(2)	-0.50
	課程對我工作上的幫助	3.68(6)	4.10(6)	-0.42

( )內數字代表該平均數在11項中的排名。

調節迴歸分析的全部結果整理如表 3 所示。從中我們有以下二點發現：首先，影響培訓滿意的主要變數爲培訓課程安排的提供與需求、及培訓課程實施的提供；影響培訓助益的主要變數爲培訓課程實施的提供、及培訓課程安排的需求；影響轉換工作意願的主要變數爲培訓課程實施的提供與需求；影響轉換工作能力的主要變數爲培訓課程實施的提供、及培訓課程安排的需求。我們可以利用以上研究結果，針對特定訓練效益變數來進行改善。

其次，所有研究假設中，只有培訓課程安排與培訓課程實施的差異間距，對於培訓滿意的影響顯著，其餘皆不顯著。我們進一步分析需求與提供的交互作用，結果如圖 7 與圖 8 所示。兩者具有相似的型態 (Patterns)，相異僅在於高需求線（圖中的粗體線）的斜率。圖 7 告訴我們，相較於低課程安排需求的學員，高課程安排需求的學員會有較高的培訓滿意。同時隨著課程安排提供的增加，所有學員（不論是低課程安排需求或高課程安排需求的學員）的培訓滿意度都會增加，但是高課程安排需求學員的培訓滿意增加程度會比較快。同樣地，圖 8 則告訴我們，相較於低課程實施需求的學員，高課程實施需求的學員會有較高的培訓滿意。同時隨著課程實施提供的增加，所有學員（不論是低課程實施需求或高課程實施需求的學員）的培訓滿意度都會增加，但是高課程實施需求學員的培訓滿意增加程度會比較快。

陸、結論

本研究運用 Kirkpatrick(1998) 的四階段評估模式，發展出可以用來衡量訓練績效的培訓滿意、培訓助益、轉換工作意願、及轉換工作能力等項目。並且以差異理論來評量學員在課程內容、課程安排、及課程實施等方面的需求與提供的感覺差異，對於訓練效益的影響。最後，再以「商業電子化人才培訓計劃」的學員為研究對象來實際驗證。

本研究對學術的貢獻，在於提供理論與實證的結果，來解釋如何衡量學員的訓練績效，並且瞭解影響這些訓練績效變數的原因。至於，對實務的貢獻則在於研究結果可以提供規劃設計人力資源訓練課程與制度時參考。後續的研究方向包括：首先，可以從事多時點觀察的縱剖面研究(Longitudinal Study) 以瞭解訓練效益變數的變化情形，同時可以更加瞭解差異理論對於訓練評估的預測能力。其次，目前的

表3：培訓提供與培訓需求對於訓練效益之迴歸分析結果

因變數	自變數	外部 培訓提供	內部 培訓需求	差異影響 顯著性a
培訓滿意(H_1)				
	培訓課程內容	+0.16 b	+0.08	不顯著
	培訓課程安排	+0.23	+0.34	顯著
	培訓課程實施	+0.20	+0.06	顯著
培訓助益(H_2)				
	培訓課程內容	+0.07	-0.24	不顯著
	培訓課程安排	+0.15	+0.48	不顯著
	培訓課程實施	+0.22	+0.30	不顯著
轉換工作意願(H_3)				
	培訓課程內容	+0.10	-0.14	不顯著
	培訓課程安排	-0.06	+0.22	不顯著
	培訓課程實施	+0.23	+0.32	不顯著
轉換工作能力(H_4)				
	培訓課程內容	+0.05	-0.21	不顯著
	培訓課程安排	+0.04	+0.45	不顯著
	培訓課程實施	+0.28	+0.16	不顯著

[a]: Interaction term adds significantly to R^2.

[b]: Beta coefficients.

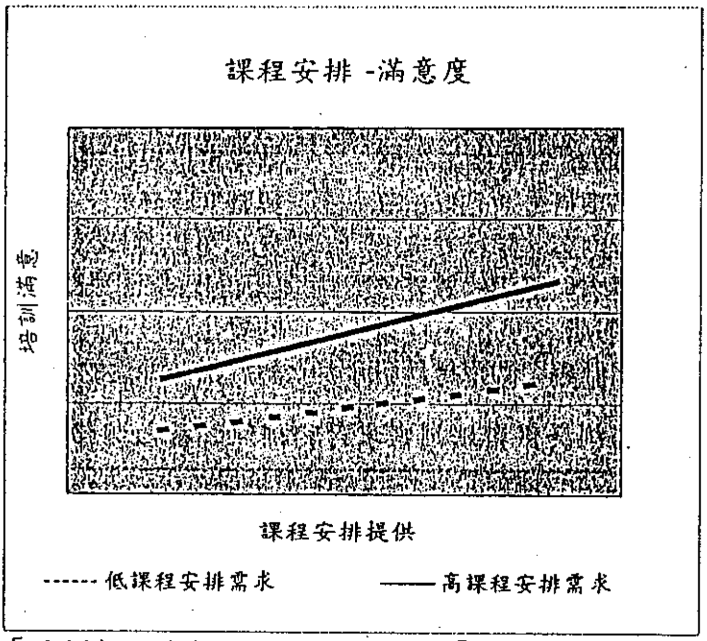


圖7：「培訓課程安排」的差異間距對「培訓滿意」之影響

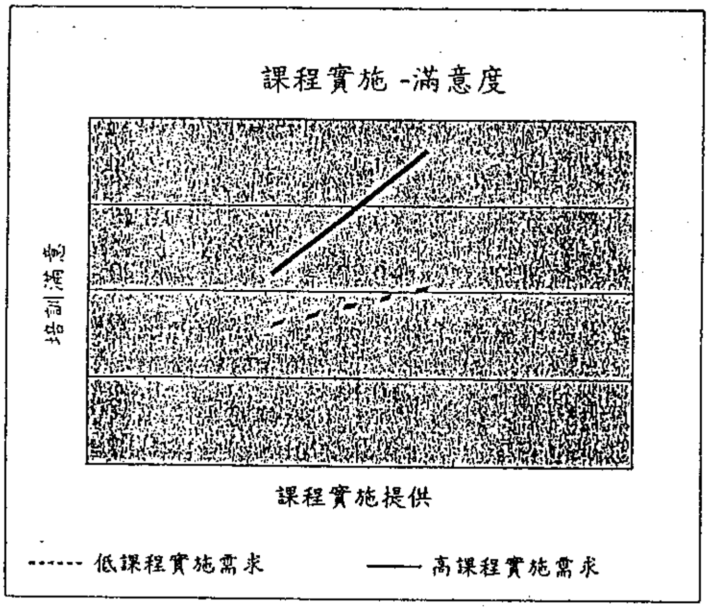


圖8：「培訓課程實施」的差異間距對「培訓滿意」之影響

研究對象皆來自單一的訓練課程，未來可以蒐集更多來自其他的訓練課程的樣本，以檢驗本研究結果的外推效度。最後，未來的學習環境可能部分會從傳統的學習環境走向 e-Learning 的環境，因此 e-Learning 下的訓練評估模式也是有潛力的研究課題。

参考文獻

1. 傅肅良，人事心理學，三版，三民書局，1985。
2. 林欽榮，人力資源管理，二版，前程出版社，1997。
3. 余品嫻，「政府部門中訓練評估之研究」，研考雙月刊 (21:1)，February 1997，84-93頁。
4. Cooper, A. C. and Artz, K. W., "Determinants of Satisfaction for Entrepreneurs," Journal of Business Venturing (10:6), November 1995, pp. 439-457.
5. Fricko, M. A. and Beehr, T. A., "A Longitudinal Investigation of Interest Congruence and Gender Concentration as Predictors of Job Satisfaction," Personnel Psychology, (45), pp. 99-117.
6. Gilles, S., "Creating Tomorrow's Consumer Bank Today," Bank Marketing (27), 1995, pp. 55-61.
7. Goldstein, I. L., Training in Organizations: Needs Assessment, Development, and Evaluation, 3rd Edition, Wadsworth, 1992.
8. Harris, M. M. and Fink, L. S., "A Field Study of Applicant Reaction to Employment Opportunities : Does The Recruiter Make a Difference?," Personnel Psychology (40), 1987, pp. 765-784.
9. Jacques, M. and Talyor, J. A. "Taking Quality as Seriously as Profits," Industrial Engineering (27), 1995, pp. 28-29.
10. Jiang, J. J. and Klein, G. A., "Discrepancy Model of IS Personnel Turnover," Working Paper, Central Florida University, March 2001.
11. Johnson, C. E. and Petrie, T. A., "The Relationship of Gender Discrepancy to Eating Disorder Attitudes and Behaviors," Sex Roles (33), September 1995, pp. 405-416.
12. Khatri, N., Fern, C. T., and Budhwar, P., "Explaining Employee Turnover in an Asian Context," Human Resource Management Journal (11:1), 2001, pp. 54-74.
13. Kirkpatrick, D. L., Evaluating Training Programs: the Four Levels, 2nd Edition, Berrett-Koehler, 1998.
14. Locke, E. A., "The Nature and Causes of Job Satisfaction," in M.D. Dunnett and L.M. Hough (Ed.), Handbook of Industrial and Organizational Psychology, Rand-McNally, 1976, pp. 1297-1349.
15. Locke, E. A., "What Is Job Satisfaction? ," Organizational Behavior and Human Performance (4), 1969, pp. 309-336.
16. Noe, R. A., "Trainees' Attributes and Attitudes: Neglected Influences on Training Effectiveness," Academy of Management Review (11:4), October 1986, pp. 736-749.
17. Piccoli, G., Ahmad, R., and Ives, B., "Web-Based Virtual Learning Environments: A Research Framework and Preliminary Assessment Effectiveness in Basic IT Skills Training," MIS Quarterly (25:4), December 2001, pp. 401 - 426.
18. Ralphs, L. T. and Stephan, E., "HRD in the Fortune 500," Training and Development Journal (40), 1986, pp. 69-76.
19. Rice, R. W., McFarlin, D. B., and Bennett, D., "Standards of Comparison

and Job Satisfaction," Journal of Applied Psychology (74:4), 1989, pp. 591-598.

20. Sarri, L. M., Johnson, T. R., McLaughlin, S. D., and Zimmerle, D. M., "A Survey of Management Training and Education Practices in U.S. Companies," Personnel Psychology, (41), 1988, pp. 731-743.

21. Thompson, J. A. and Bunderson, J. S., "Work-Nonwork Conflict and the Phenomenology of Time: Beyond the Balance Metaphor," Work and Occupations (28:1), February 2001, pp. 17-39.

附件一：問卷內容

各位親愛的學員，大家好：

首先感謝各位對於商業電子化系列課程的參與及支持。現在我們爲了瞭解各位的學習績效，並且作爲將來課程改進之依據，需要您協助塡寫下列問卷。請您務必撥出10分鐘時間，將問卷所有問題完整回答，關於您塡寫的所有內容絕對保密。我們也將在近期抽獎致贈一份精美紀念品作爲答謝。再次感謝您的協助與配合。

國立中正大學自動化研究中心
國立中正大學製商整合科技研究中心
游寶達、王俊程、黃士銘、洪新原、古政元

第一部份 請填選以下個人資料問項。

性別：□男 □女

年齡：□20歲以下 □21到30歲 □31到40歲 □41到50歲 □51歲以上

教育：□國中或以下 □高中職 □大專 □研究所或以上

婚姻：□未婚 □已婚

工作年資：□5年以下 □6到10年 □11到15年 □16到20年 □21年以上

工作現況：□不工作 □有工作 □待業中

（上個問題選「不工作」者請跳過以下二個問題；選「有工作」者，請針對現在工作，回答以下二個問題；選「待業中」者，則針對之前工作，回答以下二個問題）

職位：□一般職員 □小主管 □中階主管 □高階主管

月入：□2萬元以下 □2萬-4萬元 □4萬－6萬元 □6萬-8萬元 □8萬元以上

職業：□金融／保險業 □軍公教 □資訊業 □服務業 □醫療業
□法律相關行業 □運輸／旅遊 □娛樂／出版／傳播／行銷 □藝術
□農漁牧 □學生 □家管 □其他

第二部份 請就您個人對於下面左邊所列問題敘述的感覺，從右邊對應的五種不同程度中勾選最適合的一項。

	非常同意	同意	普通	不同意	非常不同意
我對於課程內容結構的安排相當注重。	□	□	□	□	□
我對於教材與教案的印刷與編排相當注重。	□	□	□	□	□

我對於教材與教案的提供方式相當注重。 □□□□□

我對於學習測驗的舉行方式相當注重。 □□□□□

我對於學習測驗的難易度相當注重。 □□□□□

我對於訓練場地與設施的安排相當注重。 □□□□□

我對於上課時間的分配與安排相當注重。 □□□□□

我對於講師的表達技巧相當注重。 □□□□□

我對於講師的專業知識相當注重。 □□□□□

我對於講師的教學態度相當注重。 □□□□□

我對於課程對我工作上的幫助程度相當注重。 □□□□□

訓練單位對於課程內容結構的安排相當注重。 □□□□□

訓練單位對於教材與教案的印刷與編排相當注重。 □□□□□

訓練單位對於教材與教案的提供方式相當注重。 □□□□□

訓練單位對於學習測驗的舉行方式相當注重。 □□□□□

訓練單位對於學習測驗的難易度相當注重。 □□□□□

訓練單位對於訓練場地與設施的安排相當注重。 □□□□□

訓練單位對於上課時間的分配與安排相當注重。 □□□□□

訓練單位對於講師的表達技巧相當注重。 □□□□□

訓練單位對於講師的專業知識相當注重。 □□□□□

訓練單位對於講師的教學態度相當注重。 □□□□□

訓練單位對於課程對學員工作上的幫助程度相當注重。 □□□□□

我對於課程內容結構的安排覺得滿意。 □□□□□

我對於教材與教案的印刷與編排覺得滿意。 □□□□□

我對於教材與教案的提供方式覺得滿意。 □□□□□

我對於學習測驗的舉行方式覺得滿意。 □□□□□

我對於學習測驗的難易度覺得滿意。 □□□□□

我對於訓練場地與設施的安排覺得滿意。 □□□□□

我對於上課時間的分配與安排覺得滿意。 □□□□□

我對於講師的表達技巧覺得滿意。 □□□□□

我對於講師的專業知識覺得滿意。 □□□□□

我對於講師的教學態度覺得滿意。 □□□□□

我對於課程對我工作上的幫助程度覺得滿意。 □□□□□

整體來說，我對於整個訓練課程覺得滿意。 □□□□□

由於這個訓練課程的幫助，我覺得將可爲公司的營運上貢獻更多。 □□□□□

由於這個訓練課程的幫助，我覺得將可爲公司的管理上貢獻更多。 □□□□□

由於這個訓練課程的幫助，我覺得將可爲公司的規劃上貢獻更多。 □□□□□

整體來說，由於這個訓練課程的幫助，我覺得將可爲公司貢獻更多。 □□□□□

由於這個訓練課程的幫助，我明年可能會換新工作。 □□□□□

由於這個訓練課程的幫助，我明年會積極地去找尋新的工作。 □□□□□

由於這個訓練課程的幫助，我會想要離職。 □□□□□

整體來說，由於這個訓練課程的幫助，我會想要求更好的工作。 □□□□□
由於這個訓練課程的幫助，我預期可以增加新的工作機會。 □□□□□
由於這個訓練課程的幫助，我已經獲得新的工作機會。 □□□□□
由於這個訓練課程的幫助，我已經獲得新的面談機會。 □□□□□
整體來說，由於這個訓練課程的幫助，可以增加我的工作機會。 □□□□□
整體來說，由於這個訓練課程的幫助，可以增加我轉換工作的能力。 □□□□□
對本訓練課程的建議：____
____
____。

問卷填答完畢，謝謝您的合作。倘若日後有需要一份本研究的結果，請留下您的電子郵件信箱：________________。

國立中山大學九十二學年度博士班招生考試試題

科目：資訊科技論文評述第二節【資管系】

Read the attached paper and answer the following questions. Note that the time is limited, and you should budget your time carefully. You are suggested to spend 50 minutes in reading the paper and another 50 minutes in answering the questions.

1. Please describe the work and the contributions the paper has done.
2. Please describe an application, including a scenario and an example of schema matching, to which schema matching applies.
3. Please state the comparison criteria the paper considers. Do you think they are sufficient in evaluating schema matching algorithms? Or you suggest some more it should consider. Please justify your answer.

Comparison of Schema Matching Evaluations

Hong-Hai Do, Sergey Melnik, Erhard Rahm

University of Leipzig
Augustusplatz 10-11, 04109, Leipzig, Germany
{hong, melnik, rahm}@informatik.uni-leipzig.de
dbs.uni-leipzig.de

Abstract. Recently, schema matching has found considerable interest in both research and practice. Determining matching components of database or XML schemas is needed in many applications, e.g. for E-business and data integration. Various schema matching systems have been developed to solve the problem semi-automatically. While there have been some evaluations, the overall effectiveness of currently available automatic schema matching systems is largely unclear. This is because the evaluations were conducted in diverse ways making it difficult to assess the effectiveness of each single system, let alone to compare their effectiveness. In this paper we survey recently published schema matching evaluations. For this purpose, we introduce the major criteria that influence the effectiveness of a schema matching approach and use these criteria to compare the various systems. Based on our observations, we discuss the requirements for future match implementations and evaluations.

1. Introduction

Schema matching is the task of finding semantic correspondences between elements of two schemas [11, 14, 17]. This problem needs to be solved in many applications, e.g. for data integration and XML message mapping in E-business. In today's systems, schema matching is manual; a time-consuming and tedious process which becomes increasingly impractical with a higher number of schemas (data sources, XML message formats) to be dealt with. Various systems and approaches have recently been developed to determine schema matches (semi-)automatically, e.g., Autoplex [1], Automatch [2], Clio [22, 16], COMA [7], Cupid [14], Delta [6], DIKE [19], EJX [10][1] , GLUE [9], LSD [8], MOMIS (and ARTEMIS) [3, 5], SemInt [11, 12, 13], SKAT [18], Similarity Flooding (SF) [15], and TranScm [17]. While most of them have emerged from the context of a specific application, a few approaches (Clio, COMA, Cupid, and SF), try to address the schema matching problem in a generic way that is suitable for different applications and schema languages. A taxonomy of automatic match techniques and a comparison of the match approaches followed by the various systems is provided in [20].

For identifying a solution for a particular match problem, it is important to understand which of the proposed techniques performs best, i.e., can reduce the manual work required for the match task at hand most effectively. To show the effectiveness

[1] The authors did not give a name to their system, so we refer to it in this paper using the initials of the authors' names.

of their system, the authors have usually demonstrated its application to some real-world scenarios or conducted a study using a range of schema matching tasks. Unfortunately, the system evaluations were done using diverse methodologies, metrics, and data making it difficult to assess the effectiveness of each single system, not to mention to compare their effectiveness. Furthermore, the systems are usually not publicly available making it virtually impossible to apply them to a common test problem or benchmark in order to obtain a direct quantitative comparison.

To obtain a better overview about the current state of the art in evaluating schema matching approaches, we review the recently published *evaluations* of the schema matching systems in this paper. For this purpose, we introduce and discuss the major criteria influencing the effectiveness of a schema matching approach, e.g., the chosen test problems, the design of the experiments, the metrics used to quantify the match quality and the amount of saved manual effort. We intend our criteria to be useful for future schema matching evaluations so that they can be documented better, their result be more reproducible, and a comparison between different systems and approaches be easier. For our study, we only use the information available from the publications describing the systems and their evaluation.

In Section 2, we present the criteria that we use in our study to contrast the evaluations described in the literature. In Section 3, we review the single evaluations by giving first a short description about the system being evaluated and then discussing the methodology and the result of the actual evaluation. In Section 4, we compare the evaluations by summarizing their strengths and weakness. We then present our observations concerning the current situation of the match systems as well as the challenges that future match implementations and evaluations should address. Section 5 concludes the paper.

2. Comparison criteria

To compare the evaluations of schema matching approaches we consider criteria from four different areas:

- *Input*: What kind of input data has been used (schema information, data instances, dictionaries etc.)? The simpler the test problems are and the more auxiliary information is used, the more likely the systems can achieve better effectiveness. However, the dependence on auxiliary information may also lead to increased preparation effort.
- *Output*: What information has been included in the match result (mappings between attributes or whole tables, nodes or paths etc.)? What is the correct result? The less information the systems provide as output, the lower the probability of making errors but the higher the post-processing effort may be.
- *Quality measures*: What metrics have been chosen to quantify the accuracy and completeness of the match result? Because the evaluations usually use different metrics, it is necessary to understand their behavior, i.e. how optimistic or pessimistic their quality estimation is.
- *Effort*: How much savings of manual effort are obtained and how is this quantified? What kind of manual effort has been measured, for example, pre-match effort (training of learners, dictionary preparation etc.), and post-match effort (correction and improvement of the match output)?

In the subsequent sections we elaborate on the above criteria in more detail.

2.1. Input: test problems and auxiliary information

To document the complexity of the test problems, we consider the following information about the test schemas:

- *Schema language* (relational, XML schemas, etc.): Different schema languages can exhibit different facets to be exploited by match algorithms. However, relying on language-specific facets will cause the algorithms to be confined to the particular schema type. In current evaluations, we have observed only homogeneous match tasks, i.e. matching between schemas of the same type.
- *Number of schemas and match tasks*: With a high number of different match tasks, it is more likely to achieve a realistic match behavior. Furthermore, the way the match tasks are defined can also influence the problem complexity, e.g. matching many independent schemas with each other vs. matching source schemas to a single global schema.
- *Schema information*: Most important is the number of the schema elements for which match candidates are to be determined. The bigger the input schemas are, the greater the search space for match candidates will be, which often leads to lower match quality. Furthermore, matchers exploiting specific facets will perform better and possibly outperform other matchers when such information is present or given in better quality and quantity.
- *Schema similarity*: Intuitively, a match task with schemas of the same size becomes "harder" if the similarity between them drops. Here we refer to schema similarity simply as the ratio between the number of matching elements (identified in the manually constructed match result) and the number of all elements from both input schemas [7].
- *Auxiliary information used*: Examples are dictionaries or thesauri, or the constraints that apply to certain match tasks (e.g., each source element must match at least one target element). Availability of such information can greatly improve the result quality.

2.2. Output: match result

The output of a match system is a mapping indicating which elements of the input schemas correspond to each other, i.e. match. To assess and to compare the output quality of different match systems, we need a uniform representation of the correspondences. Currently, all match prototypes determine correspondences between schema elements (element-level matches [20]) and use similarity values between 0 (strong dissimilarity) and 1 (strong similarity) to indicate the plausibility of the correspondences. However, the quality and quantity of the correspondences in a match result still depend on several orthogonal aspects:

- *Element representation*: Schema matching systems typically use a graph model for the internal representation of schemas. Hence, schema elements may either be represented by nodes or paths in the schema graphs which also impacts the representation of schema matches. Figure 1 shows a simple match problem with two small (purchase order) schemas in directed graph representation; a sample match between nodes would be *Contact↔ContactPers*. However, shared elements, such as *ContactPers* in *PO2*, exhibit different contexts, i.e. *DeliverTo* and *BillTo*, which should be considered independently. Thus, some systems return matches between node paths, e.g., *PO1.Contact↔PO2.DeliverTo.ContactPers*. Considering paths possibly leads to more elements, for which match candidates can be individually determined, and thus, possibly to more correspondences. Furthermore, the paths implicitly include valuable join information that can be utilized for generating the mapping expressions.

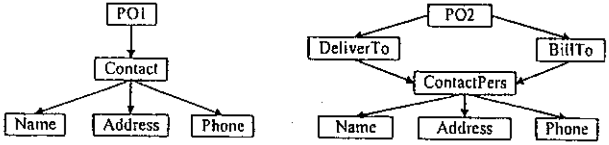


Figure 1. Schema examples for a simple match task

- *Cardinality*: An element from one schema can participate in zero, one or several match correspondences from the second input schema (*global cardinality* of 1:1, 1:n/n:1, or n:m). Moreover, within a correspondence one or more elements of the first schema may be matched with one or more elements of the second schema (*local cardinality* of 1:1, 1:n/n:1, n:m) [20]. For example, in Figure 1, *PO1.Contact* may be matched to both *PO2.DeliverTo.ContactPers* and *PO2.BillTo.ContactPers*. Grouping these two match relationships within a single correspondence, we have 1:n local cardinality. Representing them as two separate correspondences leads to 1:n global and 1:1 local cardinality. Most automatic match approaches are restricted to 1:1 local cardinality by selecting for a schema element the most similar one from the other schema as the match candidate.

2.3. Match quality measures

To provide a basis for evaluating the quality of automatic match strategies, the match task first has to be manually solved. The obtained real match result can be used as the "gold standard" to assess the quality of the result automatically determined by the match system. Comparing the automatically derived matches with the real matches results in the sets shown in Figure 2 that can be used to define quality measures for schema matching. In particular, the set of derived matches is comprised of B, the *true positives*, and C, the *false positives*. *False negatives* (A) are matches needed but not automatically identified, while false positives are matches falsely proposed by the automatic match operation. *True negatives*, D, are false matches, which have also been correctly discarded by the automatic match operation. Intuitively, both false negatives and false positives reduce the match quality.

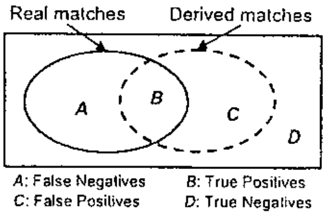


Figure 2. Comparing real matches and automatically derived matches

Based on the cardinality of these sets, two common measures, *Precision* and *Recall*, which actually originate from the information retrieval field, can be computed:

- $Precision = \frac{|B|}{|B|+|C|}$ reflects the share of real correspondences among all found ones
- $Recall = \frac{|B|}{|A|+|B|}$ specifies the share of real correspondences that is found

In the ideal case, when no false negatives and false positives are returned, we have *Precision*=*Recall*=1. However, neither *Precision* nor *Recall* alone can accurately assess the match quality. In particular, *Recall* can easily be maximized at the expense of a poor *Precision* by returning all possible correspondences, i.e. the cross product of two input schemas. On the other side, a high *Precision* can be achieved at the expense of a poor *Recall* by returning only few (correct) correspondences.

Hence it is necessary to consider both measures or a combined measure. Several combined measures have been proposed so far, in particular:

- $F\text{-}Measure(\alpha) = \frac{|B|}{(1-\alpha)*|A|+|B|+\alpha*|C|} = \frac{Precision * Recall}{(1-\alpha)* Precision + \alpha * Recall}$, which also stems from the information retrieval field [21]. The intuition behind this parametrized measure (0≤α≤1) is to allow different relative importance to be attached to *Precision* and *Recall*. In particular, *F-Measure*(α)→*Precision*, when α→1, i.e. no importance is attached to *Recall*; and *F-Measure*(α)→*Recall*, when α→0, i.e. no importance is attached to *Precision*. When *Precision* and *Recall* are considered equally important, i.e. α=0.5, we have the following combined measure:
- $F\text{-}Measure = \frac{2*|B|}{(|A|+|B|)+(|B|+|C|)} = 2*\frac{Precision * Recall}{Precision + Recall}$, which represents the harmonic mean of *Precision* and *Recall* and is the most common variant of *F-Measure*(α) in information retrieval. Currently, it is used in [2] for estimating match quality.
- $Overall = 1-\frac{|A|+|C|}{|A|+|B|} = \frac{|B|-|C|}{|A|+|B|} = Recall * \left(2-\frac{1}{Precision}\right)$, which has been introduced in [15][2] and is also used in [7]. Unlike *F-Measure*(α), *Overall* was developed specifically in the schema matching context and embodies the idea to quantify the post-match effort needed for adding false negatives and removing false positives.

To compare the behavior of *F-Measure* and *Overall*, Figure 3 shows them as functions of *Precision* and *Recall*, respectively. Apparently, *F-Measure* is much more optimistic than *Overall*. For the same *Precision* and *Recall* values, *F-Measure* is still much higher than *Overall*. Unlike the other measures, *Overall* can have negative values, if the number of the false positives exceeds the number of the true positives, i.e. *Precision*<0.5. Both combined measures reach their highest value (1.0) with *Precision*=*Recall*=1.0. In all other cases, while the value of *F-Measure* is within the range determined by *Precision* and *Recall*, *Overall* is smaller than both *Precision* and *Recall*.

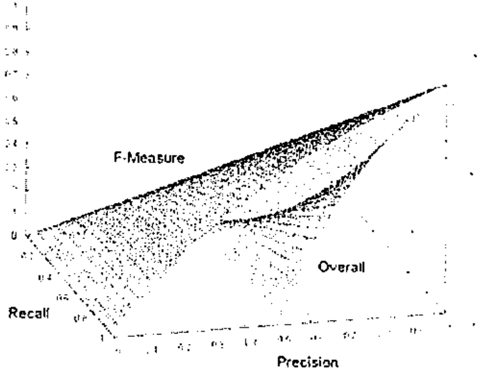


Figure 3. *F-Measure* and *Overall* as functions of *Precision* and *Recall*

[2] Here it is called *Accuracy*

2.4. Test methodology: what effort is measured and how

Given that the main purpose of automatic schema matching is to reduce the amount of manual work quantifying the user effort still required is a major requirement. However this is difficult because of many subjective aspects involved and thus a largely unsolved problem. To assess the manual effort one should consider both the *pre-match effort* required before an automatic matcher can run as well as the *post-match effort* to add the false negatives to and to remove the false positives from the final match result.

Pre-match effort includes:

- Training of the machine learning-based matchers
- Configuration of the various parameters of the match algorithms, e.g., setting different threshold and weight values
- Specification of auxiliary information, such as, domain synonyms and constraints

In fact, extensive pre-match effort may wipe out a large fraction of the labor savings obtained through the automatic matcher and therefore needs to be specified precisely. In all evaluations so far the pre-match effort has not been taken into account for determining the quality of a match system or approach.

The simple measures *Recall* and *Precision* only partially consider the post-match effort. In particular, while 1–*Recall* gives an estimate for the effort to add false negatives, 1–*Precision* can be regarded as an estimate for the effort to remove false positives. In contrast, the combined measures *F-Measure*(α) and *Overall* take both kinds of effort into account. *Overall* assumes equal effort to remove false positives and to identify false negatives although the latter may require manual searching in the input schemas. On the other hand, the parameterization of *F-Measure*(α) already allows to apply individual cost weighting schemes. However, determining that a match is correct requires extra work not considered in both *Overall* and *F-Measure*(α).

Unfortunately, the effort associated with such manual pre-match and post-match operations varies heavily with the background knowledge and cognitive abilities of users, their familiarity with tools, the usability of tools (e.g. available GUI features such as zooming, highlighting the most likely matches by thick lines, graying out the unlikely ones etc.) making it difficult to capture the cost in a general way.

Finally, the specification of the real match result depends on the individual user perception about correct and false correspondences as well as on the application context. Hence, the match quality can differ from user to user and from application to application given the same input schemas. This effect can be limited to some extent by consulting different users to obtain multiple subjective real match results [15].

3. Studies

In the following, we review the evaluations of eight different match prototypes, Autoplex, Automatch, COMA, Cupid, LSD, GLUE, SemInt, and SF. We have encountered a number of systems, which either have not been evaluated, such as Clio, DIKE, MOMIS, SKAT, and TranScm, or their evaluations have not been described with sufficient detail, such as Delta, and EJX. Those systems are not considered in our study. For each system, we shortly describe its match approach and then discuss the details of the actual evaluation. According to the taxonomy presented in [20], we briefly characterize the approaches implemented in each system by capturing

- The type of the matchers implemented (schema vs. instance level, element vs. structure level, language vs. constraint based etc.)
- The type of information exploited (e.g., schema properties, instance characteristics, and external information)
- The mechanism to combine the matchers (e.g., hybrid or composite [20, 7]).

3.1. Autoplex and Automatch

System description: Autoplex [1] and its enhancement Automatch [2] represent single-strategy schema matching approaches based on machine learning. In particular, a Naive Bayesian learner exploits instance characteristics to match attributes from a relational source schema to a previously constructed global schema. For each source attribute, both match and mismatch probability with respect to every global attribute are determined. These probabilities are normalized to sum to 1 and the match probability is returned as the similarity between the source and global attribute. The correspondences are filtered to maximize the sum of their similarity under the condition that no correspondences share a common element. The match result consists of attribute correspondences of 1:1 local and global cardinality.

Evaluation: In both Autoplex and Automatch evaluation, the global schemas were rather small, containing 15 and 9 attributes, respectively. No information about the characteristics of the involved source schemas was given. First the source schemas were matched manually to the global schema, resulting in 21 and 22 mappings in the Autoplex and Automatch evaluation, respectively. These mappings were divided into three portions of approximately equal content. The test was then carried out in three runs, each using two portions for learning and the remaining portion for matching.

The Autoplex evaluation used the quality measures *Precision* and *Recall*,[3] while for Automatch, *F-Measure* was employed. However, the measures were not determined for single experiments but for the entire evaluation: the false/true negatives and positives were counted over all match tasks. For Autoplex, they were reported separately for table and column matches. We re-compute the measures to consider all matches and obtain a *Precision* of 0.84 and *Recall* of 0.82, corresponding to an *F-Measure* of 0.82 and *Overall* of 0.66. Furthermore, the numbers of the false/true negatives and positives were rather small despite counting over multiple tasks, leading to the conclusion that the source schemas must be very small. For Automatch, the impact of different methods for sampling instance data on match quality was studied. The highest *F-Measure* reported was 0.72, so that the corresponding *Overall* must be worse.

3.2. COMA

System description: COMA [7] follows a composite approach, which provides an extensible library of different matchers and supports various ways for combining match results. Currently, the matchers exploit schema information, such as element and structural properties. Furthermore, a special matcher is provided to reuse the results from previous match operations. The combination strategies address different aspects of match processing, such as, aggregation of matcher-specific results and match candidate selection. Schemas are transformed to rooted directed acyclic graphs, on which all match algorithms operate. Each schema element is uniquely identified by

[3] Here they are called *Soundness* and *Completeness*, respectively

its complete path from the root of the schema graph to the corresponding node. COMA produces element-level matches of 1:1 local and m:n global cardinality.

Evaluation: The COMA evaluation used 5 XML schemas for purchase orders taken from www.biztalk.org. The size of the schemas ranged from 40 to 145 unique elements, i.e. paths. Ten match tasks were defined, each matching two different schemas. The similarity between the schemas was mostly only around 0.5, showing that the schemas are much different even though they are from the same domain. Some pre-match effort was needed to specify domain synonyms and abbreviations.

A comprehensive evaluation was performed with COMA to investigate the impact of different combination strategies on match quality and to compare the effectiveness of different matchers, i.e. single matchers vs. matcher combinations, with and without reuse. The entire evaluation consisted of over 12,000 test series, in each of which a different choice of matchers and combination strategies was applied. Each series in turn consisted of 10 experiments dealing with the (10) predefined match tasks. The quality measures *Precision*, *Recall*, and *Overall* were first determined for single experiments and then averaged over 10 experiments in each series (average *Precision*, etc.). Based on their quality behavior across the series, the best combination strategies were determined for the default match operation.

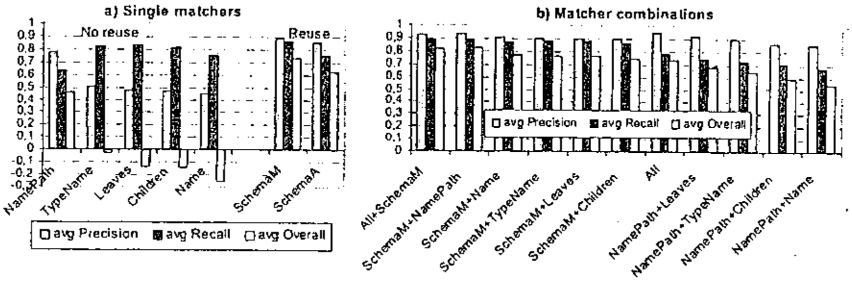


Figure 4. Match quality of COMA [7]

Figure 4a shows the quality of the single matchers, distinguished between the no-reuse and reuse-oriented ones. The reuse matchers yielded significantly better quality then the no-reuse ones. Figure 4b shows the quality of the best matcher combinations. In general, the combinations achieved much better quality than the single matchers. Furthermore, a superiority of the reuse combinations over the no-reuse ones was again observed. While the best no-reuse matcher, *All*, combining all the single no-reuse matchers, achieved average *Overall* of 0.73 (average *Precision* 0.95, average *Recall* 0.78), the best reuse combination, *All+SchemaM*, reached the best average *Overall* in the entire evaluation, 0.82 (average *Precision* 0.93, average *Recall* 0.89). These combinations also yielded the best quality for most match tasks, i.e. high stability across different match tasks. However, while optimal or close to optimal *Overall* was achieved for the smaller match tasks, *Overall* was limited to about 0.6-0.7 in larger problems. This was apparently also influenced by the moderate degree of schema similarity.

3.3. Cupid

System description: Cupid [14] represents a sophisticated hybrid match approach combining a name matcher with a structural match algorithm, which derives the similarity of elements based on the similarity of their components hereby emphasizing the name and data type similarities present at the finest level of granularity (leaf level). To address the problem of shared elements, the schema graph is converted to a tree, in which additional nodes are added to resolve the multiple relationships between a shared node and its parent nodes. Cupid returns element-level correspondences of 1:1 local and n:1 global cardinality.

Evaluation: In their evaluation, the authors compared the quality of Cupid with 2 previous systems, DIKE and MOMIS, which had not been evaluated so far. For Cupid, some pre-match effort was needed to specify domain synonyms and abbreviations. First, the systems were tested with some canonical match tasks considering very small schema fragments. Second, the systems were tested with 2 real-world XML schemas for purchase order, which is also the smallest match task in the COMA evaluation [7]. The authors then compared the systems by looking for the correspondences which could or could not be identified by a particular system. Cupid was able to identify all necessary correspondences for this match task, and thus showed a better quality than the other systems. In the entire evaluation, no quality measures were computed.

3.4. LSD and GLUE

System description: LSD [8] and its extension GLUE [9] use a composite approach tó combining different matchers. While LSD matches new data sources to a previously determined global schema, GLUE performs matching directly between the data sources. Both use machine-learning techniques for individual matchers and an automatic combination of match results. In addition to a name matcher, they use several instance-level matchers, which discover during the learning phase different characteristic instance patterns and matching rules for single elements of the target schema. The predictions of individual matchers are combined by a so-called meta-learner, which weights the predictions from a matcher according to its accuracy shown during the training phase. The match result consists of element-level correspondences with 1:1 local and n:1 global cardinality.

Evaluation: LSD was tested on 4 domains, in each of which 5 data sources were matched to a manually constructed global schema, resulting in 20 match tasks altogether. To match a particular source, 3 other sources from the same domain were used for training. The source schemas were rather small (14-48 elements), while the largest global schema had 66 attributes. GLUE was evaluated for 3 domains, in each of which two website taxonomies were matched in two different directions, i.e. A→B and B→A. The taxonomies were relatively large, containing up to 300 elements. Both systems rely on pre-match effort

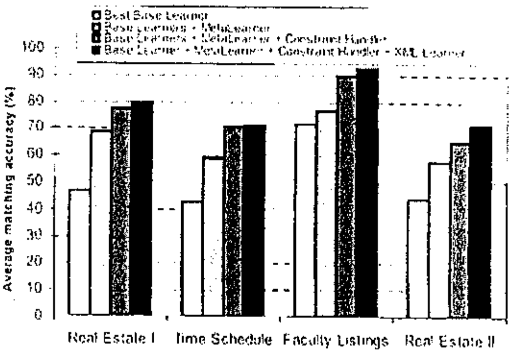


Figure 5. Match quality of LSD [8]

on the one side to train the learners, and on the other side, to specify domain synonyms and constraints.

For both LSD and GLUE, different learner combinations were evaluated. For LSD, the impact of the amount of available instance data on match quality was also studied. Match quality was estimated using a single measure, called *match accuracy*, defined as the percentage of the matchable source attributes that are matched correctly. It corresponds to *Recall* in our definition due to one single correspondence returned for each source element. Furthermore, we observe that at most a *Precision* equal to the presented *Recall* can be achieved for single match tasks; that is, if all source elements are matchable. Based on this conclusion, we can derive the highest possible *F-Measure* (=*Recall*) and *Overall* (=2**Recall*-1) for both LSD and GLUE. Figure 5 shows the quality of different learner combinations in LSD. The best quality was usually achieved when all learners were involved. In the biggest match tasks, LSD and GLUE achieved *Recall* of around 0.7, i.e. *Overall* of at most 0.4. In the case of GLUE, this quality is quite impressive considering the schema sizes involved (333 and 115 elements [9]). On average (over all domains), LSD and GLUE achieved a *Recall* of ~0.8, respectively. This corresponds to an *Overall* of at most 0.6.

3.5. Similarity Flooding (SF)

System description: SF [15] converts schemas (SQL DDL, RDF, XML) into labeled graphs and uses fix-point computation to determine correspondences of 1:1 local and m:n global cardinality between corresponding nodes of the graphs. The algorithm has been employed in a hybrid combination with a simple name matcher, which suggests an initial element-level mapping to be fed to the structural SF matcher. Unlike other schema-based match approaches, SF does not exploit terminological relationships in an external dictionary, but entirely relies on string similarity between element names. In the last step, various filters can be specified to select relevant subsets of match results produced by the structural matcher.

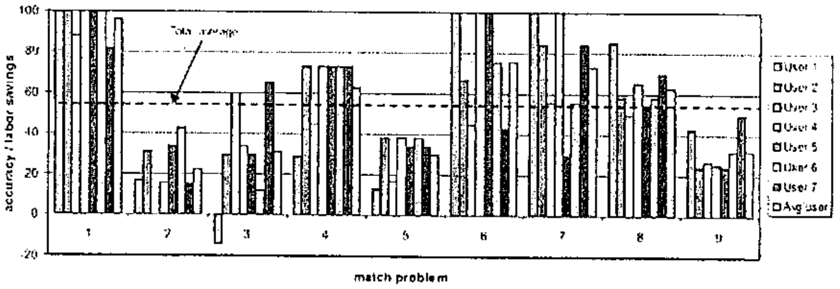


Figure 6. Match quality of Similarity Flooding Algorithm [15]

Evaluation: The SF evaluation used 9 match tasks defined from 18 schemas (XML and SQL DDL) taken from different application domains. The schemas were small with the number of elements ranging from 5 to 22, while showing a relatively high similarity to each other (0.75 on average). Seven users were asked to perform the manual match process in order to obtain subjective match results. For each match tasks, the results returned by the system were compared against all subjective results to estimate the automatic match quality, for which the *Overall* measure was used. Other experiments were also conducted to compare the effectiveness of different filters and formulas for fix-point computation, and to measure the impact of randomizing the similarities in the initial mapping on match accuracy. The best configuration was identified and used in SF. Figure 6 shows the *Overall* values achieved in the single match tasks according to the match results suggested by the single users. The average *Overall* quality over all match tasks and all users is around 0.6.

3.6. SemInt

System description: SemInt [11, 12] represents a hybrid approach exploiting both schema and instance information to identify corresponding attributes between relational schemas. The schema-level constraints, such as data type and key constraints, are derived from the DBMS catalog. Instance data are exploited to obtain further information, such as actual value distributions, numerical averages, etc. For each attribute, SemInt determines a signature consisting of values in the interval [0,1] for all involved matching criteria. The signatures are used first to cluster similar attributes from the first schema and then to find the best matching cluster for attributes from the second schema. The clustering and classification process is performed using neural networks with an automatic training, hereby limiting pre-match effort. The match result consists of clusters of similar attributes from both input schemas, leading to m:n local and global match cardinality. Figure 7 shows a sample output of SemInt. Note that each cluster may contain multiple 1:1 correspondences, which are not always correct, such as in the first two clusters.

```
(Database1.Faculty.SSN, Database1.Student.Stud_ID, Database2.Personnel.ID,
        similarity = 0.98)
(Database1.Faculty.Facu_Name, Database1.Student.Stud_Name,
        Database2.Personnel.Name, similarity = 0.92)
(Database1.Student.Tel#, Database2.Personnel.W_phone#, similarity = 0.94)
(Database1.Student.Tel#, Database2.Personnel.H_phone#, similarity = 0.95)
```

Figure 7. SemInt output: match result [12]

Evaluation: A preliminary test consisting of 3 experiments is presented in [11]. The test schemas were small with mostly less than 10 attributes. However, the quality measures for these experiments were only presented later in [12, 13]. In these small match tasks, SemInt performed very well and achieved very high *Precision* (0.9, 1.0, 1.0) and *Recall* (1.0). In [12, 13], SemInt was evaluated with two further match tasks. In the bigger match task with schemas with up to 260 attributes, SemInt surprisingly performed very well (*Precision* ~0.8, *Recall* ~0.9). But in the smaller task with schemas containing only around 40 elements, the quality dropped drastically (*Precision* 0.20, *Recall* 0.38).

On average over 5 experiments, SemInt achieved a *Precision* of 0.78 and *Recall* of 0.86. Using the *Precision* and *Recall* values presented for each experiment, we can also compute the average *F-Measure*, 0.81, and *Overall*, 0.48. On the other hand, it is necessary to take into consideration that this match quality was determined from match results of attribute clusters, each of which possibly contains multiple 1:1 correspondences. In addition to the match tasks, further tests were performed to measure the sensitivity of the single match criteria employed by SemInt [12]. The results allowed to identify a minimal subset of match criteria, which could still retain the overall effectiveness.

4. Discussion and conclusions

We first summarize the strengths and weaknesses of the single evaluations and then present our conclusions concerning future evaluations.

4.1. Comparative discussion

Table 1 gives a summary about the discussed evaluations. The test problems came from very different domains of different complexity. While a few evaluations used simple match tasks with small schemas and few correspondences to be identified (Autoplex, Automatch, SF), the remaining systems also showed high match quality for more complex real-world schemas (COMA, LSD, GLUE, SemInt). Some evaluations, such as Autoplex and Automatch, completely lack the description of their test schemas. The Cupid evaluation represents the only effort so far that managed to evaluate multiple systems on uniform test problems. Unlike other systems, Autoplex, Automatch and LSD perform matching against a previously constructed global schema.

All systems return correspondences at the element level with similarity values in the range of [0,1]. Those confined to instance-level matching, such as Autoplex, Automatch, and SemInt, can only deliver correspondences at the finest level of granularity (attributes). In all systems, except for SemInt, correspondences are of 1:1 local cardinality, providing a common basis for determining match quality.

Only the SF evaluation took into account the subjectivity of the user perception about required match correspondences. Unlike other approaches, SemInt and SF do not require any manual pre-match effort. In several evaluations, e.g. COMA, LSD, GLUE, SemInt and SF, different system configurations were tested by varying match parameters on the same match tasks in order to measure the impact of the parameters on match quality. Those results have provided valuable insights for improving and developing new match algorithms.

Usually, the quality measures were computed for single match experiments. Exceptions are Cupid with no quality measure computed, and Autoplex, Automatch with quality measures mixing the match results of several experiments in a way that does not allow us to assess the quality for individual match tasks. Whenever possible, we tried to translate the quality measures considered in an evaluation to others not considered so that one can get an impression about the actual meaning of the measures. Still, the computed quality measures cannot be used to directly compare the effectiveness of the systems because of the great heterogeneity in other evaluation criteria. Only exploiting schema information, COMA seems quite successful, while the LSD/GLUE approach is promising for utilizing instance data.

4.2. Conclusions

The evaluations have been conducted in so different ways that it is impossible to directly compare their results. While the considered match problems were mostly simple, many techniques have proved to be quite powerful such as exploiting element and structure properties (Cupid, SF, COMA), and utilizing instance data, e.g., by Bayesian and Whirl learners (LSD/GLUE) or neural networks (SemInt). Moreover, the combined use of several approaches within composite match systems proved to be very successful (COMA, LSD/GLUE). On the other side, there are still unexploited opportunities, e.g. in the use of large-scale dictionaries and standard taxonomies and increased reuse of

Table 1. Summary of the evaluations

	Autoplex & -match	COMA	Cupid	LSD & GLUE	SemInt	SF
References	[1] and [2]	[7]	[14]	[8] and [9]	[11, 12, 13]	[15]
Test problems						
Tested schema types	relational	XML	XML	XML	relational	XML, relational
#Schemas / #Match tasks	15/21 & 15/22	5/10	2/1	24/20 & 3/6	10/5	18/9
Min/Max/Avg schema size	-	40/145/77	40/54//47	14/66/- & 34/333/143	6/260/57	5/22/12
Min/Max/Avg schema similarity	-	0.43/0.8/0.58	-	-	-	0.46/0.94/0.75
Match result representation						
Matches	element-level correspondences with similarity value in range [0,1]					
Element repr.	node (attr.)	path	path	node	node (attr.)	node
Local/global cardinality	1:1/1:1	1:1/m:n	1:1/n:1	1:1/n:1	m:n/m:n (attr. cluster)	1:1/m:n
Quality measures and test methodology						
Employed quality measures	Precision, Recall & F-Measure	Precision, Recall, Overall	none	Recall	Precision, Recall	Overall
Subjectivity	1 user					7 users
Studied impact on match quality	Automatch: methods for sampling instance data	matchers, combination, reuse, schema characteristics	none	learner combinations, LSD: amount of data listings	constraints (discriminators)	filters, fix-point formulas, randomizing initial sim.
Pre-match effort	training	specifying domain synonyms	specifying domain synonyms	training, specifying domain synonyms, constraints	none	none
Best average match quality						
Prec./Recall	0.84/0.82	0.93/0.89	-	~0.8/0.8	0.78/0.86	-
F-Measure	0.82 & 0.72	0.90	-	~0.8	0.81	-
Overall	0.66	0.82	-	~0.6	0.48	~0.6
Evaluation highlights						
		Big schemas, Systematic evaluation	Comparative evaluation of 3 systems	Big schemas	Big schemas, No pre-match effort	User subjectivity, No pre-match effort

previous match results (COMA). Future match systems should integrate those techniques within a composite framework to achieve maximal flexibility.

Future evaluations should address the following issues:

- *Better conception for reproducibility*: To allow an objective interpretation and easy comparison of match quality between different systems and approaches, future evaluations should be conceived and documented more carefully, if possible, including the criteria that we identified in this paper.
- *Input factors – test schemas and system parameters*: All evaluations have shown that match quality degrades with bigger schemas. Hence, future systems should be evaluated with schemas of more realistic size, e.g. several hundreds of elements. Besides the characteristics of the test schemas, the various input parameters of each system can also influence the match quality in different ways. However, their impact has rarely been investigated in a comprehensive way, thus potentially missing opportunities for improvement and tuning. Similarly, previous evaluations

typically reported only some peak values w.r.t. some quality measure so that the overall match quality for a wider range of configurations remained open.

- *Output factors – match results and quality measures*: Instead of determining only one match candidate per schema element, future systems could suggest multiple, i.e. *top-K*, match candidates for each schema element. This can make it easier for the user to determine the final match result in cases where the first candidate is not correct. In this sense, a top-K match prediction may already be counted as correct if the required match candidate is among the proposed choices.
  Previous studies used a variety of different quality measures with limited expressiveness thus preventing a qualitative comparison between systems. To improve the situation and to consider precision, recall and the degree of post-match effort we recommend the use of combined measures such as *Overall* in future evaluations. However, further user studies are required to quantify the different effort needed for finding missing matches, removing false positives, and verifying the correct results. Another limitation of current quality measures is that they do not consider the pre-match effort and the hardness of match problems.

Ultimately, a *schema matching benchmark* seems very helpful to better compare the effectiveness of different match systems by clearly defining all input and output factors for a uniform evaluation. In addition to the test schemas, the benchmark should also specify the use of all auxiliary information in a precise way since otherwise any hard-to-detect correspondences could be built into a synonym table to facilitate matching. Because of the extreme degree of heterogeneity of real-world applications, the benchmark should not strive for general applicability but focus on a specific application domain, e.g., a certain type of E-business. Alternatively, a benchmark can focus on determining the effectiveness of match systems with respect to specific match capabilities, such as name, structural, instance-based and reuse-oriented matching. Currently we are investigating how such benchmarks could be generated.

5. Summary

Schema matching is a basic problem in many database and data integration applications. We observe a substantial research and development effort in order to provide semi-automatic solutions aiding the user in this time-consuming task. So far, many systems have been developed and several of them evaluated to show their effectiveness. However, the way the systems have been tested varies to a great extent from evaluation to evaluation. Thus it is difficult to interpret and compare the match quality presented for each system.

We proposed a catalog of criteria for documenting the evaluations of schema matching systems. In particular, we discussed various aspects that contribute to the match quality obtained as the result of an evaluation. We then used our criteria and the information available in the literature to review several previous evaluations. Based on the observed strengths and weaknesses, we discussed the problems that future system implementations and evaluations should address. We hope that the criteria that we identified provide a useful framework for conducting and describing future evaluations.

Acknowledgements

We thank Phil Bernstein and AnHai Doan for many useful comments. This work is supported by DFG grant BIZ 6/1-1.

References

1. Berlin, J., A. Motro: Autoplex: Automated Discovery of Content for Virtual Databases. CoopIS 2001, 108–122
2. Berlin, J., A. Motro: Database Schema Matching Using Machine Learning with Feature Selection. CAiSE 2002
3. Bergamaschi, S., S. Castano, M. Vincini, D. Beneventano: Semantic integration of heterogeneous information sources. Data & Knowledge Engineering 36: 3, 215–249, 2001
4. Bright, M. W., A.R. Hurson, S. Pakzad: Automated Resolution of Semantic Heterogeneity in Multidatabase. ACM Trans. Database Systems 19: 2, 212–253, 1994
5. Castano, S., V. De Antonellis: A Schema Analysis and Reconciliation Tool Environment. IDEAS 1999, 53–62
6. Clifton, C., E. Housman, E., A. Rosenthal: Experience with a Combined Approach to Attribute-Matching Across Heterogeneous Databases. IFIP 2.6 Working Conf. Database Semantics 1996
7. Do, H.H., E. Rahm: COMA – A System for Flexible Combination of Schema Matching Approach. VLDB 2002
8. Doan, A.H., P. Domingos, A. Halevy: Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. SIGMOD 2001
9. Doan, A.H., J. Madhavan, P. Domingos, A. Halevy: Learning to Map between Ontologies on the Semantic Web. WWW 2002
10. Embley, D.W., D. Jackman, L. Xu: Multifaceted Exploitation of Metadata for Attribute Match Discovery in Information Integration. WIIW 2001
11. Li, W.S., C. Clifton: Semantic Integration in Heterogeneous Databases Using Neural Networks. VLDB 1994
12. Li, W.S., C. Clifton: SemInt: A Tool for Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Network. Data and Knowledge Engineering 33: 1, 49–84, 2000
13. Li, W.S., C. Clifton, S.Y. Liu: Database Integration Using Neural Networks: Implementation and Experiences. Knowledge and Information Systems 2: 1, 2000
14. Madhavan, J., P.A. Bernstein, E. Rahm: Generic Schema Matching with Cupid. VLDB 2001
15. Melnik, S., H. Garcia-Molina, E. Rahm: Similarity Flooding: A Versatile Graph Matching Algo-rithm. ICDE 2002
16. Miller, R.J. et al. The Clio Project: Managing Heterogeneity. SIGMOD Record 30:1: 78–83, 2001
17. Milo, T., S. Zohar: Using Schema Matching to Simplify Heterogeneous Data Translation. VLDB 1998, 122–133
18. Mitra, P., G. Wiederhold, J. Jannink: Semi-automatic Integration of Knowledge Sources. Fusion 1999
19. Palopoli, L., G. Terracina, D. Ursino: The System DIKE: Towards the Semi-Automatic Synthesis of Cooperative Information Systems and Data Warehouses. ADBIS-DASFAA 2000, 108–117
20. Rahm, E., P.A. Bernstein: A Survey of Approaches to Automatic Schema Matching. VLDB Journal 10: 4, 2001
21. Van Rijsbergen, C. J.: Information Retrieval. 2nd edition, 1979, London, Butterworths.
22. Yan, L.L., R.J. Miller, L.M. Haas, R. Fagin. Data-Driven Understanding and Refinement of Schema Mappings. SIGMOD, 2001