Sampling issues and management solutions in internet-based market researches

Maria V. Ciasullo, Giuseppe Festa
Università degli Studi di Salerno, Dipartimento di Studi e Ricerche Aziendali (Management & Information Technology)
Via Giovanni Paolo II – 132, 84084 Fisciano (SA)
mciasullo@unisa.it, gfesta@unisa.it

Questo lavoro è frutto di un impegno comune. Tuttavia, i paragrafi 1, 2 e 6 sono da attribuirsi a Maria V. Ciasullo, mentre i paragrafi 3, 4 e 5 sono da attribuirsi a Giuseppe Festa.

Abstract. *In the wide context of internet-based market researches, this paper focuses on those ones carried out on net users. An analysis of the advantages, but also disadvantages, related to online market researches shows that a crucial problem is the 'vagueness' of the representativeness of sample surveys, up to affecting the credibility of the online survey. Therefore, the authors have developed a theoretical / practical framework that integrates contributions from marketing, statistics and informatics, following a logical-methodological path oriented to give minor uncertainty degrees to online surveys. This model recovers useful considerations from inferential statistics and proposes also some solutions for managing internet samples as virtual communities. The conceptual paper aims to contribute to an advancement in the field of online market researches, shifting the focus about the validity / reliability of the investigation from the mere software or statistical tool to a wider analytical internet marketing strategy.*

Keywords*: e-research, analytical internet marketing, online sampling.*

1. Introduzione

L'orientamento *customer based* [Valdani e Busacca, 2000], requisito inderogabile per la sopravvivenza e il successo delle organizzazioni, caratterizza in modo sempre più preponderante l'agire delle imprese dei nostri giorni. In tale logica, diviene fondamentale un'approfondita conoscenza delle esigenze di chi acquista, del suo modo di interpretare la realtà e dei meccanismi attraverso i quali operano i processi di scelta. Parallelamente, un'economia sempre più globalizzata, caratterizzata da una rapida trasferibilità delle informazioni, ha contribuito a delineare non solo un cliente più attento e complesso, ma anche nuovi stili di consumo basati sulla capacità dei beni di esprimere la cultura, il gusto, lo stile di ciascun individuo.

I continui cambiamenti associati a una maturazione dei mercati e delle relative dinamiche innovative di acquisto impongono alle imprese un costante impegno nell'analisi e nella comprensione dei fenomeni anzidetti. Diviene pertanto comprensibile come, nell'ambito degli studi di marketing management, sia stata posta particolare attenzione alla funzione analitica del marketing, tesa ad analizzare la struttura del mercato e l'evoluzione della domanda, al fine di individuare i fattori più significativi che possano influenzare i comportamenti d'acquisto.

Le ricerche di mercato, generalmente intese quali indagini sistematiche e obiettive finalizzate alla raccolta di elementi rilevanti per successivamente elaborare e implementare strategie di interazione con i clienti e con gli altri attori del sistema del valore dell'impresa, assolvono alla finalità di osservare mercati e delineare scenari. È quindi evidente come le stesse costituiscano *driver* fondamentali nell'agire dell'impresa, laddove la competitività di quest'ultima si basi sulla capacità di conquistare e alimentare costantemente il sostegno a una domanda matura, esigente e innovativa.

Nella complessità di tale contesto s'inserisce la valenza delle tecnologie ad alta intensità connettiva (prevalentemente, se non esclusivamente, *internet-based*), capace di contribuire in modo significativo alla ricerca sociale e, nello specifico, alle ricerche di mercato. Internet si presenta infatti quale strumento "naturalmente" idoneo a raccogliere informazioni in maniera mirata, approfondita e metodologicamente corretta sulla domanda [Kiang et al., 2000], sia nei termini delle ricerche di mercato sia ancor prima tramite opportune operazioni di *marketing intelligence* [Kotler, 2001]. In tale prospettiva, si sottolinea un'importante proprietà della rete, ossia la sua capacità di rendere i processi comunicativi sostanzialmente indipendenti dai vincoli spazio-temporali: è evidente, infatti, come tale proprietà si dimostri estremamente valorizzante per la ricerca online, in quanto consente il superamento di una serie di barriere fisiche che possono esistere tra il ricercatore e i soggetti sui quali viene condotta la rilevazione.

Inoltre, come evidenziato in letteratura [Cantone et al., 2006; Vescovi, 2007], le nuove tecnologie dell'informazione e della comunicazione, in particolare quelle che si basano su standard aperti e universali, assumono un ruolo fondamentale nel miglioramento della gestione del sistema delle relazioni intra e inter-impresa e dei sottostanti processi di creazione di valore, facilitando altresì la fase di avvicinamento a culture e mercati internazionali. Per tutte queste ragioni, è evidente come l'efficacia delle ricerche di mercato richieda un continuo aggiornamento che si combina inevitabilmente all'evoluzione *internet-based* dei sistemi informativi di marketing e quindi, nella prospettiva di questo lavoro, alle strategie di internet marketing analitico in senso lato.

2. Le ricerche di mercato internet-based: natura, tassonomia e problematica

La combinazione tra marketing e statistica è tradizionalmente alla base del sistema di competenze che regola il mondo delle ricerche di mercato [Bassi, 2008]; a voler essere più precisi, in particolare, il marketing, in questo particolare contesto, si serve della statistica inferenziale prevalentemente nella fase del campionamento, allo scopo di estendere alla popolazione le osservazioni sul campione (*sample*). Nel tempo, a questo binomio di competenze ha finito per accodarsi una terza disciplina, ossia l'informatica, che, in ragione dei vantaggi conseguibili prevalentemente in termini di tempi e costi, ha contribuito a potenziare notevolmente l'interazione tra le due competenze anzidette, al punto che oggi, quasi senza timore di smentita, il sistema delle ricerche di mercato è per l'appunto governato dalla sintesi di queste tre discipline (marketing, statistica e informatica).

L'avvento dell'internet, inoltre, non ha soltanto generato una nuova serie di strumenti informatici a disposizione del ricercatore, ma, in ragione del suo successo "sociale", ha finito per consegnare al mondo delle ricerche di mercato anche un nuovo polmone d'indagine. In tal modo, pertanto, diventa possibile distinguere tra le ricerche di mercato svolte grazie alle nuove soluzioni in rete (e parleremo propriamente di ricerche di mercato "con" internet) e le ricerche di mercato svolte presso gli utenti della rete (e parleremo propriamente di ricerche di mercato "su" internet): le due categorie insieme, evidentemente, costituiscono la famiglia allargata delle *ricerche di mercato internet-based*, che, in altre parole, in un modo o in un altro hanno a che fare con l'internet.

Nell'ambito del presente lavoro, in particolare, non ci soffermeremo sulla prima categoria ("con"), in primo luogo perché esclusivamente strumentale e in secondo luogo perché già analizzata in un precedente lavoro [Festa, 2001b]. In questa sede, invece, si vuole ragionare della seconda categoria ("su"), identificandone la principale problematica di riferimento (ossia la rappresentatività campionaria) e proponendone, se possibile, un'analisi teorica, che tenga in conto il contributo integrato proprio del marketing, della statistica e dell'informatica.

Si ritiene, anzitutto, che un corretto approccio metodologico alla tematica delle ricerche di mercato *internet-based* non possa prescindere dall'impostazione delle ricerche di mercato tradizionali, verificando se contestualmente la relativa "cassetta degli attrezzi" (in termini di tecniche e strumenti) sia inapplicabile (perché non adattabile all'ambiente internet), applicabile con indispensabile adattamento (dovendo necessariamente operare in modalità informatica) o, infine, applicabile in chiave innovativa (sfruttando le notevoli possibilità dell'ambiente internet). A muovere questa classificazione, che parte da un mero ragionamento epistemologico [Festa, 2001a], sono prevalentemente due fattori: le caratteristiche "tecnologiche" della piattaforma internet e le caratteristiche "sociali" della popolazione internet, che per brevità si analizzano congiuntamente grazie a una rapida panoramica dei vantaggi e degli svantaggi collegati alle ricerche di mercato online (senza voler ovviamente approfondire la tematica, che di per sé sarebbe vastissima).

Per quanto riguarda alcuni dei principali vantaggi, si pensa, prevalentemente, all'efficienza e all'economicità, in termini di risparmi sui tempi (data la velocità di elaborazione delle attuali attrezzature informatiche, hardware e software) e sui costi (non in senso assoluto, ma ovviamente se paragonati ai costi normali di una ricerca di mercato tradizionale), arrivando a una sorta di consacrazione dell'internet quale canale più efficiente per la raccolta dei dati [Weible e Wallace, 1998]. Tuttavia, esistono anche vantaggi più significativamente qualitativi: le potenzialità dell'elaborazione informatica, infatti, non sono da considerare soltanto in riferimento ai dati immagazzinati, ma anche e soprattutto in riferimento alle modalità di immagazzinamento, capaci di realizzare un'esperienza più attraente per l'intervistato (grazie alla multimedialità), una maggiore indipendenza dei tempi e degli spazi di rilevazione (e, quindi, maggiore comodità e conseguenziale precisione di rilevazione, sia per l'intervistato sia per l'intervistatore), una maggiore indipendenza dal contesto (personale e/o situazionale: si pensi a disabili, ammalati, reclusi, zone di guerra, zone di epidemia, ecc.), un abbattimento delle barriere psicologiche e socioculturali (facendo leva sulla sensazione di assenza della personalità fisica dell'interlocutore) e così via [Schmitz, 2004; Wright, 2005].

Il complessivo maggior senso di comfort, inoltre, produce in genere risultati più ampi e profondi, soprattutto nelle ricerche qualitative, che, fin d'ora e come si suggerirà per altre motivazioni anche in seguito, sembrano costituire uno scenario di approfondimento altamente interessante per le ricerche di mercato *internet-based* [Andreani e Conchon, 2001b], richiedendo contestualmente nuove competenze al ricercatore (si pensi, per esempio, alle abilità richieste al moderatore di un gruppo online). In generale, infatti, la ricerca qualitativa aspira a ottenere una profonda comprensione del comportamento del consumatore, attraverso un'analisi approfondita delle aspirazioni e delle sensazioni associate per esempio a una categoria, un brand, un'immagine [Mariampolski, 2001; Marbach, 2010]. Al di là della rappresentatività statistica (indispensabile per le tecniche quantitative), le tecniche qualitative presentano il vantaggio di essere applicabili anche nei casi in cui l'informazione sia difficile da rilevare, complessa da analizzare o di problematica interpretazione [Andreani, 1998; Tissier-Desbordes, 1998; Andreani e Conchon, 2001a] proprio perché conferiscono maggiore libertà espressiva ai soggetti intervistati [Prandelli e Verona, 2006]. In questo modo, grazie a internet, l'intervistato diventa completamente "padrone" dell'intervista, senza che l'intervistatore possa giocare da "manipolatore". Se così non fosse e dovesse accorgersene, infatti, all'intervistato basterebbe un clic per andarsene o, peggio ancora, per imbrogliare l'intervistatore.

Al progettista della ricerca qualitativa su internet e al relativo intervistatore, pertanto, serviranno competenze tecnologiche, metodologiche e soprattutto socio-relazionali perfettamente in linea con l'ambiente internet. Anche per questi motivi, in particolare, le ricerche di mercato più interessanti su internet sono a nostro avviso proprio le ricerche qualitative, perché possono far emergere informazioni che in altri contesti difficilmente si può essere

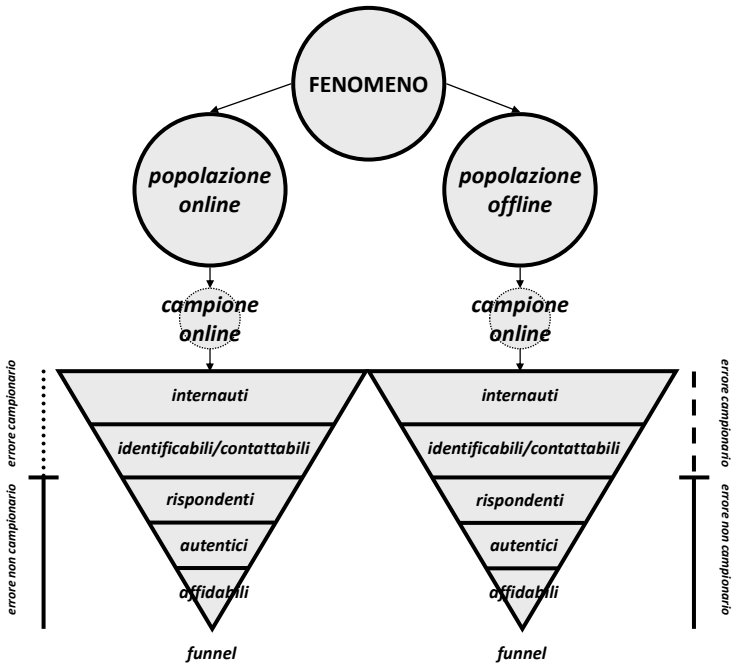
disposti a confessare (anche ricorrendo alla distorsione complessivamente generabile con il linguaggio para-verbale e non-verbale).

Esistono naturalmente anche diversi svantaggi collegati alle ricerche di mercato online, che sembrano, in ogni caso, ampiamente compensati dai relativi vantaggi, già commentati in precedenza: si fa riferimento, per esempio, alla maggiore o minore familiarità dell'utente con il "linguaggio" della tastiera e del mouse (si pensi al monitoraggio di una chat), alla perdita d'informazione sul contesto (si pensi a un questionario online a cui si risponde in ufficio oppure da casa, dovendo evidentemente "sottostare" a regole di comportamento diverse in ambienti diversi) e all'impossibilità di raccogliere segnali non-verbali e para-verbali (e forse anche verbali: si pensi, per esempio, a modi di esprimersi fisiologicamente diversi nella scrittura elettronica rispetto all'uso della penna o ancor di più della voce, ecc.). Tuttavia, come evidenziato da molteplici studi [Howard et al., 2001; Andrews et al., 2003; McDonald e Stewart, 2003; Di Fraia, 2004; Wright, 2005] la principale mancanza delle ricerche su internet è legata alla "vaghezza" della rappresentatività dell'indagine campionaria, che finisce per diventare un problema di notevole importanza, al punto da condizionare la stessa credibilità dell'indagine online.

Avendo ampiamente verificato in letteratura la consapevolezza di tale lacuna, la principale finalità di questo paper di ricerca concettuale è quella di elaborare un *framework* teorico / pratico che possa contribuire a meglio delimitare il perimetro del problema della rappresentatività campionaria su internet, proponendo altresì specifiche soluzioni gestionali, derivanti da una complessiva strategia di internet marketing analitico. In tale logica, si ritiene di poter far emergere alcuni specifici spunti di riflessione che potranno essere oggetto di ulteriori ricerche in materia.

3. Alcune caratteristiche delle indagini campionarie su internet

L'analisi della problematica in oggetto sembra rappresentabile per successive "differenze": rispetto a un fenomeno da osservare, infatti, non tutta la popolazione di riferimento potrebbe essere attiva su internet (*internauti*); degli utenti internet che appartengano (anche) alla popolazione di riferimento, non tutti potrebbero essere *identificabili / contattabili* dal ricercatore ai fini del campionamento; anche tra quelli contattati, non tutti potrebbero decidere di rispondere (*rispondenti*); e tra questi, infine, non tutti potrebbero rispondere con sincerità (*autentici*) o con cura (*affidabili*), in ragione della distanza spazio-temporale "fisica" dell'indagine e del senso di "irresponsabilità" socialmente diffuso in rete. Si pensi, per esempio, a quanti utenti internet si avvalgano di servizi gratuiti (e-mail, video sharing, freeware, ecc.) alla cui registrazione, obbligatoriamente richiesta dal provider a fini prevalentemente commerciali, abbiano fornito dati completamente inattendibili o persino inventati, pur di superare l'ostacolo inevitabile della registrazione e usufruire del servizio internet. Questi ostacoli permettono di costruire un vero e proprio "imbuto" (*funnel*), che associamo graficamente a tale problematica, come rappresentato in Figura 1.



(Figura 1) Errori campionari e non campionari nell'impiego di campioni online

Dalla rappresentazione grafica si evince facilmente che le ricerche di mercato su internet, riferite a popolazioni online oppure offline, presentano ovviamente identiche problematiche per quanto riguarda l'errore non campionario (ossia errori dell'intervistatore, dell'intervistato, del ricercatore, ecc.), in cui bisogna gestire i soggetti rispondenti, autentici e affidabili, ma soltanto "a valle della base campionaria" (ossia l'insieme dei soggetti identificabili /

contattabili su internet). In precedenza, invece, si pone quell'evidente questione di rappresentatività, che è alla base del ragionamento sviluppato nel presente lavoro, ossia la possibile (pressoché certa) discrepanza tra la popolazione offline e il campione online, circostanza che ovviamente contribuisce a rendere più complesso l'errore campionario (graficamente rappresentato con un tratteggio più intenso).

Bisogna infatti specificare che l'errore campionario in senso stretto è in linea di principio identico per entrambi i campioni (e questa osservazione costituirà proprio una delle basi per la soluzione metodologica di seguito proposta), perché deriva dall'applicazione del teorema del limite centrale, secondo il quale all'aumentare della numerosità campionaria aumenta la rappresentabilità del fenomeno (o, più precisamente, all'aumentare della numerosità campionaria la distribuzione campionaria del fenomeno tende alla distribuzione normale standardizzata, consentendo stime attendibili dopo aver definito errore e rappresentatività) [Natale, 2004]. Tuttavia, è chiaro che un campione online è "fisiologicamente" ereditario di una popolazione online, mentre nel caso di un campione online impiegato per rappresentare una popolazione offline tale ereditarietà è inevitabilmente spuria.

Poiché proprio queste sembrano le fondamentali problematiche riguardanti le ricerche online, a nostro avviso non soltanto non è possibile arrivare a una tassonomia esaustiva dei dispositivi impiegabili nelle ricerche di mercato *internet-based*, perché gli strumenti internet sono in continua evoluzione (tecnologica e sociale) e soprattutto è in continua evoluzione la loro combinazione, ma, dal punto di vista del marketing, tale sforzo potrebbe essere tutto sommato nemmeno dirimente. Piuttosto, si può tentare una rappresentazione della problematica campionaria e della relativa soluzione seguendo uno schema concettuale, cominciando dall'impostazione logico-metodologica della ricerca di mercato e proseguendo con una visione sinottica delle strumentazioni e, soprattutto, delle tecniche disponibili e/o attuabili.

In tal senso, nel voler seguire rigorosamente l'impostazione classica delle ricerche di mercato, ferma restando la necessaria elasticità rispetto alle diverse situazioni professionali, è doveroso attingere in primo luogo a una solida base concettuale su cui innescare successivi approfondimenti, legati alla natura dell'ambiente internet. Pertanto, un approccio corretto (o, almeno, metodologicamente non scorretto) alle ricerche di mercato *internet-based* è proprio identificabile nel tradizionale processo della ricerca di mercato, articolabile in otto fasi [Barile e Metallo, 2002]:

a) definizione del problema di marketing;
b) identificazione della finalità della ricerca;
c) formulazione degli obiettivi dell'indagine;
d) progettazione dell'intervento (in termini di qualità, tempi e costi);
e) valutazione economica delle azioni da intraprendere;
f) raccolta dei dati (primari e/o secondari, da fonti interne e/o esterne);
g) elaborazione e analisi dei dati (con eventuali proposte di applicazione);
h) relazione finale (corredata, come si vedrà successivamente, delle "credenziali").

Si ribadisce come questa impostazione non sia chiaramente obbligatoria, ma è ovviamente facile apprenderne i vantaggi in termini di schematicità. Proseguendo nella strutturazione, la fase più delicata del processo, ossia la "progettazione dell'intervento", può essere a sua volta articolata in ulteriori otto sotto-fasi, come può evincersi dal seguente percorso [ibidem]:

1) formulazione degli obiettivi operativi della ricerca;
2) progettazione dell'indagine in termini concreti e fattuali;
3) progettazione del campione;
4) selezione e strutturazione degli strumenti di rilevazione;
5) gestione delle risorse umane impegnate nell'indagine;
6) programmazione temporale delle attività da svolgere;
7) pianificazione delle tabulazioni e delle codificazioni dei risultati;
8) stima dei costi operativi.

Si evidenzia come questa sotto-articolazione sia perfettamente inserita nel processo logico-metodologico della ricerca di mercato. Si noti, infatti, come la prima (*formulazione degli obiettivi operativi della ricerca*) e l'ultima sotto-fase (*stima dei costi operativi*) della "progettazione dell'intervento" siano fisiologicamente collegate alla fase precedente ("formulazione degli obiettivi dell'indagine") e a quella successiva ("valutazione economica delle azioni da intraprendere").

Anche soltanto con una rapida scorsa delle fasi e sotto-fasi così esposte, risulta ben chiara la presenza di diverse "zone" concettuali caratterizzate da identica problematicità, sia nelle ricerche *earth-based* sia in quelle *internet-based* (si pensi, per esempio, all'analisi del problema di marketing). Altre "zone", invece, presentano situazioni da contestualizzare in termini di strumentazioni (si pensi, per esempio, al questionario online e non più cartaceo), ma capaci di realizzare i vantaggi più evidenti collegati alle ricerche di mercato online. Esiste, infine, una "zona" caratterizzata da specifiche problematicità attinenti le ricerche di mercato *internet-based*, tra le quali la più rilevante è senza dubbio la rappresentatività, come già si anticipava in precedenza e come si è in questo modo verificato anche metodologicamente.

Osservazione (grazie al *funnel*) e metodologia (grazie alla *literature review*) permettono quindi di determinare come, in definitiva, la principale problematica relativa alle ricerche *internet-based* risieda nell'ancora limitata rappresentatività dell'indagine. Per cercare di superare tale limite, nella visione proposta in questo lavoro si ritiene indispensabile un supporto da parte del marketing alla statistica inferenziale, contrariamente a quanto avviene tradizionalmente nelle ricerche di mercato, in cui il contributo statistico è funzionale all'analisi ed eventualmente alla soluzione del problema di marketing [Molteni e Troilo, 2003].

Infatti, la matematica e la statistica, che costituiscono le discipline maggiormente contribuenti allo studio della rappresentatività del campione, non potranno che correttamente discendere da regole GIGO ("garbage in, garbage out") [Seglin, 1994] e, in un certo senso, disinteressarsi della qualità del dato di partenza: proprio su tale questione, invece, dovrà soffermarsi il ricercatore, il quale ben intuisce le potenzialità dell'internet e si adopera per il superamento dei contestuali limiti. Al momento, infatti, la problematica in oggetto si riferisce alla rappresentatività del campione online rispetto a fenomeni (anche) offline, mentre è evidente che un'indagine su un fenomeno online potrebbe avere seri problemi di rappresentatività se svolta offline, ossia presso intervistati non normalmente fruitori dei servizi internet oggetto dell'indagine.

Paradossalmente, pertanto, si dimostra che le ricerche online, pur patendo profondamente il "funnel", non possono essere considerate inattendibili a priori, perché, per esempio, è epistemologicamente sbagliato indagare mediante tecniche tradizionali un utente, anche soltanto in sede di *concept test*, in merito all'esperienza di navigazione in un *virtual store* multimediale, magari anche tridimensionale. Alcune proposte in tal senso sono sviluppate nella parte finale del paper, in cui si ragiona sulla metodologia di raccolta, all'interno di *internet sample* concepiti come comunità virtuali [Hagel e Armstrong, 1997], del dato "autentico" e "affidabile" rispetto al problema di marketing in questione.

4. Riflessioni sulle possibili soluzioni al problema della rappresentatività online

Si proverà, di seguito, a ragionare sulle principali caratteristiche del campione adeguatamente rappresentativo, esponendo genericamente alcune motivazioni statistiche e successivamente soffermandosi su approfondimenti gestionali, che, ovviamente, più sono in sintonia con lo spirito di questo lavoro.

In generale, la rappresentatività del campione (*n*) rispetto alla popolazione (*N*) dovrebbe essere conseguita, supposta la correttezza nel calcolo della numerosità campionaria, tramite il campionamento probabilistico, in primo luogo tramite il campionamento casuale semplice [Molteni e Troilo, 2003]: infatti, soltanto un campione casuale semplice (ossia con probabilità di campionamento nota e uguale per tutti i membri della base campionaria) garantisce l'assenza dell'errore di selezione (che certamente figura, invece, nel campionamento non probabilistico, ossia per convenienza, giudizio e quota). Proprio l'azzeramento dell'errore di selezione, ricavabile dal campionamento casuale semplice, dovrebbe assicurare, in concomitanza come si diceva di un'adeguata numerosità, anche assenza di distorsione.

Nel campionamento non probabilistico, peraltro, il ricercatore non è ignaro della possibile distorsione, ma anzi la usa a proprio vantaggio, ritenendo che quel consapevole errore di selezione gli consenta una rappresentatività contestualmente meno infelice, anche se, ovviamente, non può saperlo *ex ante*. Dall'estrazione casuale semplice, infatti, potrebbe venir fuori proprio il campione non probabilistico, che sarebbe più rappresentativo, perché casuale, rispetto allo stesso (identico) campione non probabilistico: in tale possibilità risiede il paradosso del campionamento casuale semplice, che trova la propria forza non nel risultato, ma nel procedimento: del resto, anche un campione casuale (forse soprattutto un campione casuale) può produrre un valore estremo.

Il campionamento, come si anticipava in precedenza, si basa sul teorema del limite centrale, secondo il quale l'aumento della numerosità campionaria rende la distribuzione campionaria sempre più simile a una distribuzione normale standardizzata, consentendo pertanto di calcolare la numerosità campionaria necessaria per sopportare un dato margine di errore (Es_x) e un dato intervallo di confidenza (z). Infatti, possiamo ricorrere alle seguenti formule (cfr. Tabella 1) per il calcolo della numerosità campionaria (è infinita, per convenzione statistica, la popolazione superiore a 100.000 elementi; è binomiale il fenomeno rappresentabile in due sole modalità, mentre diversamente è continuo) [Barile e Metallo, 2002].

	Popolazione finita	**Popolazione infinita**
Fenomeno continuo	$n = N z^2 s^2 / [(N - 1) E^2 s_x + z^2 s^2]$	$n = (z s / Es_x)^2$
Fenomeno binomiale	$n = N z^2 p (1 - p) / [(N - 1) E^2 s_x + z^2 p (1 - p)]$	$n = p (1 - p) (z / Es_x)^2$

(Tabella 1) Calcolo della numerosità campionaria (n) rispetto alla numerosità della popolazione (N)

È del tutto evidente, pertanto, che per la precisione della stima non conta tanto la frazione di campionamento (f = n / N) quanto la numerosità campionaria, proprio in forza del teorema del limite centrale. Queste considerazioni, essenziali nella teoria dei campioni, permettono di affermare che la questione della rappresentatività su internet, per certi versi, ispirati più alla pratica dell'indagine che alla scientificità della metodologia, possa trascurare un primo stadio del *funnel*, ossia la non coincidenza della popolazione osservata con l'utenza internet, purché si riesca a costruire un campione online sufficientemente numeroso e ovviamente aggiornato rispetto alla mortalità statistica.

In pratica, il problema più rilevante della rappresentatività online si riferisce a quel numero di elementi della base campionaria che concorra a formare l'adeguata numerosità campionaria calcolata, risultando una caratteristica aggiuntiva e non penalizzante a priori il fatto che tali elementi della popolazione offline siano anche utenti attivi online. Invece, reclutare esclusivamente utenti attivi online per un campione di una popolazione offline potrebbe per certi versi essere assimilato a un campionamento non probabilistico [Andreani e Conchon, 2001a], che avvenga per giudizio o per convenienza (eventualmente, anche per quota: si pensi, per esempio, a un'indagine sulla soddisfazione dei clienti di una banca che sia attiva con filiali tradizionali e con servizi internet).

Nel campionamento non probabilistico, infatti, non ha alcun valore il concetto di errore standard, alla base, assieme alla confidenza, del calcolo della numerosità campionaria. A servirsi del campionamento non probabilistico sono prevalentemente le ricerche di mercato qualitative, che, anche per questo motivo, sembrano essere le indagini meglio praticabili tramite ricerche di mercato online, non avendo per natura pretesa di descrizione, ma soltanto di esplorazione.

Si è ragionato, pertanto, del come la rappresentatività del campione online non debba essere collegata tanto alla grandezza dell'utenza internet rispetto alla popolazione, quanto piuttosto alla numerosità dell'utenza internet indagata. In tal senso, di conseguenza, potrebbe non essere necessario estrarre un campione online da una popolazione online, corrispondente in tutto o in parte alla popolazione offline, perché potrebbe bastare il positivo riscontro delle caratteristiche della popolazione presso il campione, purché adeguatamente numeroso. Si seguirebbe, in tal modo, un procedimento non di estrazione (dall'alto verso il basso), ma di astrazione (dal basso verso l'alto).

È fondamentale, quindi, mitigare le impurità collegate all'indagine con la trasparenza delle relative credenziali: caratteristiche della popolazione, tipologia di campionamento, numerosità campionaria, percentuale di errore ammissibile e confidenza della stima. A queste indicazioni statistiche bisognerebbe correttamente aggiungere tutte le altre informazioni ritenute di una qualche influenza nell'estrapolazione del risultato (una sorta di nota integrativa alla ricerca): considerazioni sulla formazione degli intervistatori, sulla formulazione delle domande, sulla catalogazione delle risposte e così via, facendo anche attenzione alla corretta implementazione di sistemi per l'eliminazione o almeno la mitigazione dell'errore non campionario online, con particolare riferimento agli stadi del "funnel" relativi a *rispondenti*, *autentici* e *affidabili*.

Tali credenziali rappresentano non soltanto la "cartina di tornasole" per l'interpretazione dei risultati dell'indagine, ma anche e soprattutto la corretta chiave di lettura per la valutazione delle informazioni ottenute, diventando le stesse credenziali, pertanto, informazioni per il processo decisionale. Infatti, la naturale distorsione statistica del campionamento, dovuta alla fisiologica distanza tra universo e campione, non può essere messa in discussione: come vale per le ricerche *earth-based*, vale certamente anche per quelle *internet-based* (la natura socio-economica è perennemente cangiante e anche gli utenti internet, ovviamente, sono persone).

Pertanto, la problematica relativa alla rappresentatività della popolazione internet rispetto ai fenomeni offline, che costituisce il principale *focus* di questo studio, sembra poter essere risolta facendo tesoro delle riflessioni logico-metodologiche esposte in precedenza (anche se ovviamente con l'adeguata considerazione delle caratteristiche volta per volta contestuali del fenomeno). Tramite una complessiva strategia di internet marketing analitico, finalizzata a costruire, monitorare e gestire gli *internet sample* come se fossero comunità virtuali, diventa inoltre possibile ragionare sui meccanismi utili a garantire, a valle, la "pulizia" del dato in ingresso, risolvendo in questo modo gli ulteriori limiti rappresentati dal *funnel*, con una possibile proposta operativa che viene di seguito analizzata.

5. Alcune proposte operative per la gestione di un internet sample

In base al ragionamento esposto nel presente lavoro, sembra possibile delineare alcune fondamentali caratteristiche, soprattutto metodologiche (*strategiche*) e applicative (*operative*), non soltanto per la corretta progettazione, ma anche per la corretta gestione di un *internet sample* per le ricerche online. In tal senso, sembra particolarmente interessante pensare a una sorta di comunità virtuale, da costruire con abnegazione e attenzione, spostando la prospettiva di riferimento dalle ricerche di mercato in senso stretto a una più complessiva strategia di internet marketing, basata (anche quando analitico) sul binomio visibilità e attrattività [Metallo e Festa, 2003].

In altre parole, sembra paradossalmente non dirimente porsi problemi eccessivi sulla validità e sull'attendibilità delle ricerche di mercato online in quanto tali, perché la principale problematica di riferimento sembra riguardare invece la numerosità della popolazione online, da cui estrapolare la base campionaria e successivamente il campione per l'indagine. Quest'ultimo, se adeguatamente numeroso, potrebbe essere sufficiente in termini di rappresentatività almeno da un punto di vista empirico: da un punto di vista metodologico, invece, tale possibilità potrebbe anche non essere mai praticabile, non essendo sostanzialmente possibile procedere su internet a un campionamento casuale semplice che sia certamente rappresentativo di una popolazione offline [Schmitz, 2004].

In questa direzione, il *funnel* deve essere continuamente aggiustato e mitigato, nel senso che bisogna insistere costantemente non soltanto sulla progettazione statistica del campione, ma anche e soprattutto sulla complessiva correttezza operativa nella gestione del campione, attivando una serie di meccanismi di continua autenticazione dell'intervistato, orientata a garantire identità e sincerità nella rilevazione.

Questo continuo imbrigliamento, al contempo, potrebbe essere causa di eccezioni sulla complessiva spontaneità e naturalezza dell'intervistato nonché sulla sua predisposizione all'intervista (la buona lena nel rispondere, per dire, potrebbe essere fiaccata da continue autenticazioni), ma appare per le motivazioni finora esposte una delle poche strade al momento perseguibili per incrementare la complessiva attendibilità dell'indagine online. I passaggi di autenticazione dovrebbero essere molteplici e potrebbero essere schematizzati come segue:

- *iscrizione*: l'utente deve in primo luogo iscriversi alla comunità virtuale (il ricercatore, quindi, deve fin dall'inizio dell'indagine immaginare il campione online di una qualsiasi ricerca online come una comunità virtuale, potendone ricavare anche un senso di maggior "appartenenza" dell'intervistato al buon esito della ricerca), compilando un *form* di registrazione, in base al quale il ricercatore dovrebbe desumere una prima serie di dati per procedere alla classificazione dell'utente in base ai principali criteri di segmentazione dell'intera comunità. Gli strati in questione non sono statici e nemmeno dinamici: sono semplicemente "virtuali" (allo stesso modo in cui, in un database, le *query* possono essere considerate delle tabelle virtuali), potendo sfruttare la capacità del sistema informatico sottostante alla comunità virtuale di ricavare, grazie a opportune operazioni di filtraggio, la segmentazione più utile in quel preciso momento (tramite comuni applicazioni OLAP). A tale registrazione il sistema assegna, su scelta dell'utente e conformemente alle politiche della comunità, un *account*, costituito da *username* e *password*, che d'ora in avanti saranno la chiave di accesso al sistema, con relativa autenticazione, autorizzazione e responsabilità;
- *accreditamento*: in sede di svolgimento della singola rilevazione, all'utente dovrebbe essere richiesta non soltanto l'autenticazione (tramite *account*), ma anche un successivo livello di autorizzazione, identificabile p.e. in un *captcha* (sulla pagina web di accesso alla rilevazione) e/o in una *password* (nell'e-mail di "convocazione all'indagine" spedita all'utente). *Captcha* è l'acronimo dell'espressione "completely automated public Turing test to tell computers and humans apart" ("test di Turing pubblico e completamente automatico per distinguere tra macchine e umani"): per confermare l'uso del servizio internet da parte di un individuo e non di una macchina, viene proposta a video un'immagine con dei caratteri, che l'utente è chiamato a digitare in un'apposita casella di testo per la verifica. I caratteri nell'immagine sono artatamente distorti nell'apparenza, ma perfettamente comprensibili a occhio nudo e, pertanto, sono interpretabili da un individuo e non da un macchina (aumentando, si spera, il livello di attenzione dell'intervistato). In questo modo, il sistema si garantisce rispetto alla possibilità che l'utente (o lo stesso sistema informatico sottostante alla comunità virtuale) sia inconsapevolmente infettato da un *malware*, che potrebbe quindi accedere alla comunità virtuale, ma non alla rilevazione, in cui potrebbe generare risposte inutili;
- *verifica*: infine, per avere l'ulteriore certezza che a essere oggetto della rilevazione sia proprio l'utente, diventa inevitabile costringere quest'ultimo a fornire adeguate risposte a determinate domande, alle quali egli soltanto può rispondere. Tali domande dovrebbero essere presentate durante la rilevazione in modalità *random*, così da "costringere" quello specifico utente a essere "fisicamente" presente nell'interazione con il sistema. È da evidenziare che l'autenticazione degli utenti rappresenta un elemento centrale delle infrastrutture di sicurezza informatica e le tecniche basate sulla conoscenza sono attualmente il metodo usato più frequentemente per l'autenticazione dell'utente [Ellison et al., 2000; Dhamija e Perrig, 2000). Nello specifico, attraverso la *knowledge based authentication* (KBA) l'utente, per poter essere riconosciuto come tale, deve rispondere correttamente a una domanda (standard) alla quale ha già risposto in precedenza e della quale, ovviamente, il sistema informatico conserva la memoria della relativa risposta (generando, probabilmente, anche una maggior empatia con l'utente, il quale dovrebbe / potrebbe vedersi non come un "numero" della batteria degli intervistati, ma come un soggetto al quale è richiesto di interagire in maniera univocamente individuale). Tale tecnica è ampiamente usata per ripristinare accessi a servizi internet nel caso in cui l'utente abbia dimenticato la password: nel caso di risposta corretta alla domanda standard, il sistema invia un'e-mail all'indirizzo precedentemente lasciato dall'utente, ricordandogli username e password (o generandone una nuova). Alcune domande, infine, potrebbero essere concepite anche come vere e proprie "domande di controllo", allo scopo di recuperare la maggiore affidabilità possibile in merito alle risposte dell'intervistato [De Luca, 2006].

Come si accennava in precedenza, questo sistema di continua autenticazione potrebbe rallentare la rilevazione e renderla persino scomoda per l'utente, ma al momento queste sembrano le soluzioni gestionali più incisive nell'efficacia e diffuse nella pratica. Tali riflessioni, naturalmente, riguardano anche altri meccanismi di accesso, basati non su quello che l'utente "sa" (come analizzato finora), ma su quello che l'utente "ha" (per esempio, una smart card, un token, ecc.) e/o su quello che l'utente "è" (scansione biometrica) [Teti e Festa, 2009]: anche in questi casi, tuttavia, nulla vieta di sospettare che l'utente possa accedere al sistema una prima volta con le proprie credenziali, per lasciare svolgere a qualcun altro la rilevazione in un momento successivo.

Con la continua tecnica dell'autenticazione (*iscrizione-accreditamento-verifica*) si fa quindi riferimento a una pratica di costante verifica della correttezza comportamentale dell'utente (eliminando o mitigando le problematiche relative agli stadi relativi a *rispondenti*, *autentici* e *affidabili* del "funnel"); come in tutte le attività di controllo, si finirà per essere meno produttivi da un punto di vista operativo, assicurandosi però una migliore qualità dell'indagine. Potrebbe anche accadere, a dire il vero, che parte degli utenti selezionati abbandoni l'indagine o persino l'*internet sample* gestito come comunità virtuale: a questo punto, bisognerebbe chiedersi a quale compromesso si possa arrivare tra quantità e qualità del campione. Anche per tale motivo sembrano maggiormente adeguate all'ambiente internet le ricerche qualitative, meno sensibili alla problematica del campionamento e alla stessa numerosità campionaria, perché principalmente finalizzate, come già ampiamente evidenziato, a meccanismi di esplorazione e non di descrizione.

6. Risultati, implicazioni e conclusioni

L'evoluzione dell'informatica e, in particolare, l'introduzione, la diffusione e lo sviluppo dell'internet hanno avuto, continuano ad avere e avranno sempre di più in futuro uno straordinario impatto sui quotidiani processi economici e sociali. Queste innovazioni hanno naturalmente interessato anche il mondo dei sistemi informativi aziendali (anzi, soprattutto questo mondo), provocando radicali trasformazioni nella progettazione, nel funzionamento e nello sviluppo delle procedure fisiche (processi organizzativi) e logiche (flussi informativi) chiamate a svolgere specifiche attività d'impresa.

Una delle discipline maggiormente coinvolte in tale cambiamento è quella delle ricerche di mercato, che si trova a poter usare internet come strumento aggiuntivo rispetto agli "attrezzi" tradizionali (ricerche "con" internet) o persino come bacino aggiuntivo / sostitutivo / innovativo di soggetti da sottoporre a indagine (ricerche "su" internet). Tra queste due possibilità, oltre all'inevitabile esigenza di procedere a un'adeguata contestualizzazione delle nuove strumentazioni informatiche, le maggiori problematiche di studio derivano dalle ricerche "su" internet, soprattutto in considerazione della distanza (sociale, culturale, economica, ecc.) che potrebbe esserci (che, anzi, costantemente esiste) tra popolazione offline e popolazione online, con ovvie ricadute sull'apprezzamento della rappresentatività del campione online rispetto alla popolazione offline.

In questo paper, chiaramente orientato a uno studio di natura concettuale, si è pertanto proceduto in primo luogo a una rappresentazione logica delle principali problematiche di campionamento nelle ricerche online, arrivando a sviluppare un *framework* teorico / pratico ("funnel") che riesca a tenere complessivamente conto degli errori, campionari e non campionari, che potrebbero metodologicamente emergere nel collegamento di rappresentatività (da una parte) tra popolazione online e campione online e (dall'altra) tra popolazione offline e campione online. Grazie a tale schematizzazione, ricorrendo a continue riflessioni in contaminazione tra marketing, statistica e informatica, è stato possibile proporre alcune soluzioni empiriche per la gestione della rappresentatività online, più di statistica "sostanziale" nel caso dell'errore campionario e più di marketing "applicato" nel caso dell'errore non campionario.

Tali soluzioni, in ogni caso, sono utilmente proponibili soltanto se perseguite nella prospettiva dell'*internet sample* concepito come comunità virtuale, in cui i soggetti sottoposti alla rilevazione siano considerati dal ricercatore (e soprattutto da sé stessi) membri di uno specifico gruppo attivo online, anche se soltanto per la durata dell'indagine. Diventa fondamentale, pertanto, immaginare una strategia di internet marketing analitico per alimentare e mantenere il campione / comunità, facendo ricorso alla cangiante combinazione tra visibilità e attrattività.

Le implicazioni scientifiche del presente lavoro sembrano riguardare soprattutto le possibili evoluzioni dell'*humus* alla base della disciplina delle ricerche di mercato, ossia la tradizionale combinazione tra marketing, statistica e (ormai) informatica, individuando quindi in tale mix di competenze un probabile oggetto di ulteriori studi, sia in termini metodologici che applicativi. In ragione della pervasività dell'internet, infatti, gli approfondimenti sulle ricerche di mercato dovranno sempre di più confrontarsi con un'evoluzione sociale della Rete che già espone (e sempre di più esporrà in futuro) alla necessità / opportunità di trovare soluzioni ragionevoli (per l'oggi) e/o lungimiranti (per il futuro), come nello spirito delle riflessioni e delle proposte sviluppate in questo studio, che sembrano pertanto fornire interessanti spunti applicativi soprattutto in termini di implicazioni operative (e quindi per chi sia chiamato da subito a svolgere ricerche online).

Non bisogna infatti dimenticare che il mondo delle ricerche di mercato online è ancora ai suoi primordi (come lo sono in generale tutte le ambientazioni *internet-based*) per quanto riguarda i valori, i principi e soprattutto le regole di funzionamento (si pensi alla netiquette, alla privacy, ai social network, ecc.): tale caratteristica, assieme alla vertiginosa velocità di sviluppo delle strumentazioni internet e delle relative possibilità d'uso, che comprime esponenzialmente la durata dell'*internet year*, rende già obsoleta o almeno rischiosa qualsiasi soluzione, soprattutto se empirica (ed è questo un probabile limite della ricerca concettuale che si è presentata). Per tale motivo, è sempre fondamentale recuperare, contestualizzare e sviluppare un rigoroso impianto logico-metodologico, com'è avvenuto in questo studio, che ha attinto dalla "tradizionale" metodologia delle ricerche di mercato, successivamente proponendo, prima in merito alla prospettiva dell'internet e successivamente in merito alla prospettiva del campionamento su internet, soluzioni sostanzialmente gestionali (se si vuole, di "buon senso", come normalmente avviene nell'economia d'impresa), non limitandosi alla sola analisi dell'utilità del singolo strumento statistico o informatico, ma mescolando continuamente osservazioni e rappresentando costantemente applicazioni rispondenti a una complessiva strategia di internet marketing analitico.

Bibliografia

Andreani J.C., Conchon F., Gli studi qualitativi in Internet, Micro & Macro Marketing, 1, 2001a, 65-74.

Andreani J.C., Conchon F., Les études qualitatives en marketing, Paris: ESCP-EAP, Les Cahiers de Recherche, 2001b, 01-150.

Andreani J.C., L'interview qualitative en marketing, Revue Française du Marketing, 1988, 3-4.

Andrews D., Nonnecke B., Preece J., Electronic survey methodology: A case study in reaching hard-to-involve Internet users, International Journal of Human-Computer Interaction, 16, 2, 2003, 185-210.

Barile S., Metallo G., Le ricerche di mercato. Aspetti metodologici e applicativi, Giappichelli, Torino, 2002.

Bassi F., Analisi di mercato. Strumenti statistici per le decisioni di marketing, Carocci, Roma, 2008.

Cantone L., Calvosa P., Testa P., Tecnologie ad alta intensità connettiva, relazioni di marketing e processi di creazione di valore per i clienti nei mercati BtB, Mercati & Competitività, 2006, 3, 1.38.

De Luca A., Le ricerche di mercato. Guida pratica e metodologica, Angeli, Milano, 2006.

Dhamija R., Perrig A., D´ej`a Vu: A User Study Using Images for Authentication, Proceedings of the 9th USENIX Security Symposium Denver, Colorado, USA, August 14-17, 2000, 1-15.

Di Fraia G., Validità e attendibilità delle ricerche on line, in Di Fraia G. (a cura di), e-Research. Internet per la ricerca sociale e di mercato, Laterza, Roma-Bari, 2004.

Ellison C., Hall C., Milbert R., Schneier B., Protecting secret keys with personal entropy, Future Generation Computer Systems, 16, 2000, pp. 311-318.

Festa G., Il ruolo dei nuovi strumenti informatici nel sistema informativo di marketing, Esperienze d'impresa, 2, 2001b, pp. 67-114.

Festa G., Digital Business English - Glossario ragionato di linguistica d'impresa per la new economy, Edisud, 2001a.

Hagel J., Armstrong A., Net gain, Harvard Business School Press, Boston, 1997.

Howard P., Rainie L., Jones S., Days and nights on the Internet: The impact of a diffusing technology, American Behavioral Scientist, 45, 3, 2001, 383–404.

Kiang M.Y., Raghu T.S., Huei-Min Shang K., Marketing on the Internet — who can benefit from an online marketing approach?, Decision Support Systems, 27, 2000, 383-393.

Kotler P., Marketing Management, Pearson Education Italia, Milano, 2001.

Marbach G., Ricerche per il Marketing, Utet, Torino, 2010.

Mariampolski H., Qualitative Market Research, Sage, London, 2001.

Martone D., Furlan R., Online Market Research Tecniche e metodologia delle ricerche di mercato tramite Interne, Franco Angeli, Milano, 2007.

McDonald H., Stewart A., A comparison of online and postal data collection methods, Marketing Intelligence & Planning, 21, 2, 2003, 85-95.

Melody Y. Kiang M.Y., Raghu T.S., Huei-Min Shang K., Marketing on the Internet — who can benefit from an online marketing approach?, Decision Support Systems, 2000, 27, 383–393.

Metallo G., Festa G., La progettazione dei portali web nell'ottica del customer service, in Barile S., Metallo G. (a cura di) Soluzioni problematiche d'impresa. Riflessioni e modalità risolutive, Edizioni Culturali Internazionali, Roma, 2003.

Molteni L., Troilo G., Ricerche di marketing, McGraw-Hill, Milano, 2003.

Natale P., Il sondaggio, Laterza, Bari, 2004.

Prandelli E., Verona G., Marketing in rete, oltre Internet verso il nuovo Marketing, McGraw Hill, Milano, 2006.

Schmitz N., Rappresentatività e campionamenti nelle survey on line, in Di Fraia G. (a cura di), e-Research. Internet per la ricerca sociale e di mercato, Laterza, Roma-Bari, 2004.

Seglin J.L., Marketing (la guida al), McGraw-Hill, Milano, 1994, pag. 37.

Teti A., Festa G., Sistemi informativi per la sanità, Apogeo, Milano, 2009.

Tissier-Desbordes E., Les études qualitatives dans un monde post-moderne, Revue Française du Marketing, 1998, 3-4.

Valdani E., Busacca B., Customer Based View: dai principi alle azioni, Proc. de Convegno Le tendenze del Marketing in Europa), Università Ca' Foscari Venezia, 2000, 1-24.

Vescovi T., Il Marketing e la rete. La gestione integrata del web nel business. Comunicazione, e-commerce, sales management, business to business. Il Sole 24Ore, Milano 2007.

Weible R., Wallace J., Cyber research: the impact of the Internet on data collection, Market Research, 1998, 10, 3, 19-31.

Wright K.B., Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services, Journal of Computer-Mediated Communication, 10, 3, 2005.

Document Image Retrieval by Layout Analysis

G. Pirlo[1] , M. Chimienti[2], M. Dassisti[3], D. Impedovo[4], A. Galiano[4]

[1] *Dipartimento di Informatica, Università degli Studi di Bari "A. Moro", via Orabona 4, 70125-Bari, Italy*
[2] *Laboratorio Kad3, C.da Baione, 70043 Monopoli (BA), Italy*
[3] *Dip. Meccanica, Management e Matematica, Politecnico di Bari, viale Japigia 182, 70126 - Bari, Italy*
[4] *Dyrecta Lab, Via V. Simplicio 45, 70014 Conversano (BA) - Italy*

(corresponding author: giuseppe.pirlo@uniba.it)

A new layout-based document image retrieval system is presented in this paper. The system is specifically designed for commercial form retrieval and uses mathematical morphology to extract structural components from the document image. Document layout description is performed by the Radon Transform whereas Dynamic Time Warping is used for matching. The experimental results have been carried out on both real and simulated data sets. They demonstrate the effectiveness of the proposed approach and their robustness with respect to different classes of commercial forms and shifted/rotated document images.

Index Terms *— Document management, Document Image Retrieval, Mathematic Morphology, Radon Transform, Dynamic Time Warping.*

1. Introduction

The increasing number of documents available in databases and digital libraries makes document retrieval a critical task of current document management systems. Traditional document retrieval systems – based on set-theoretic, algebraic and probabilistic models - require a document to be present in text form and the querying method is based on a specific textual content in the document [Doermann, 1998; Manning et al., 2009]. Whatever the model used, text-based document retrieval systems require a document in text form, since the search for similar documents is based on comparing the textual contents. As a consequence, a preliminary stage of image to text conversion by an Optical Character Recognizer (OCR) is required when a document is in image form. OCR is a time-consuming error-prone process, specifically in the

case of multi-lingual/multi-font documents and poor-quality document images [Marukawa et al., 1997; Taghva et al., 1996; Lorpesti, 1996], as discussed in comprehensive surveys on this topic [Doermann, 1998; Mitra and Chaudhuri, 2000].

Along with the spreading of multimedia documents, it is useful to search a document on the basis of its structure and not only on the basis of its textual content. In such cases, methods adopted for document retrieval use feature vectors in which each feature is extracted from a specific region of the document image. For instance, some researchers used a static zoning strategy for document image decomposition to extact a fixed-size feature vector from the document image. In this approach, a regular grid is superimposed to the document image in order to extract regional characteristics [Tzacheva et al., 2002]. In another approach, a hierarchical zoning strategy was proposed to overcome the problem of optimal grid selection, in order to face with the treatment of set of documents of different characteristics [Duygulu and Atalay, 2002]. A system that extracts text lines and describes the layout by means of relationships between pairs of these lines was also discussed in the literature [Huang et al., 2005], whereas some researchers used Brick Wall Coding Features (BWC) features to represent bounding boxes of the words [Erol et al., 2008]. Although the features are scale invariant and robust to slight perspective distortion, the accuracy of their system is very low. In addition the method does not work correctly when documents are written in languages such as Japanese and Chinese, in which words are not separated. Several approaches can also be combined to identify a document, such as barcode, micro optical patterns, encoding hidden information, paper fingerprint, character recognition, local features and RFID. Owing to utilize SIFT. Unfortunately, the retrieval process is time consuming and requires special equipment [Liu and Liao, 2011].

In this paper a Layout-based Document Image Retrieval (LDR) system is presented, that is specifically devoted to commercial form processing, such as invoices, waybills, receipts, etc., in which layout is strongly characterized by a grid-structure. In fact, in these particular cases, traditional document-image approaches are not effective since they are not able to describe documents on the basis of the grid-based structure. In the first step the system uses a technique based on mathematical morphology for removing textual components from the document image and for extracting the grid-based structure in the document layout. Subsequently Radon Transform is used to obtain the feature vector characterizing the specific grid structure of the document. Dynamic Time Warping (DTW) is finally adopted to perform document matching.

The paper is organized as follow. The architecture of the system is presented in Section 2. Section 3 describes the preprocessing phase, which uses operators of mathematical morphology. The feature extraction phase is presented in Section 4. In Section 5 the matching process based on DTW is discussed while the decision combination process is illustrated in Section 6. The experimental results are reported and discussed in Section 7. Section 8

presents the conclusion of the work and highlights some directions for further research.

2. The Radon Transform for Layout-based Document Image Retrieval

The LDR system presented in this paper is based on three main phases: Acquisition and Preprocessing; Feature Extraction; Matching. After document image acquisition, the document is preprocessed and transformed by Radon Transform. The features extracted are then stored in the reference database in the enrollment stage. In the running stage, an unknown document is first scanned and preprocessed, successively the features are extracted compared to the those stored into the database. The matching module performs matching by Dynamic Time Warping (DTW) and outputs the ranked list of similar documents. More precisely, the input document is acquired as a standard 256 gray-level – 100dpi PDF file. Figures 1 shows an input document concerning a real invoice.

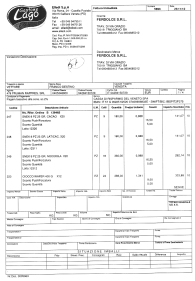


Figure 1. Input document image I=I(x,y).

Successively, after noise removal, document is resampled to 100 dpi and grid-based structure is extracted by mathematical morphology [Serra, 1982]. More precisely, let I=I(x,y) be the document image ($1 \le x \le X$, $1 \le y \le Y$) and let be

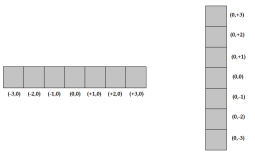
- B_{hor} the horizontal structure element defined as (see Figure 2a):

$B=\{(-s,0), \ldots, (-1,0), (0,0), (1,0), \ldots , (s,0)\}$;

- B_{ver} the horizontal structure element defined as (see Figure 2b):

$B=\{(0,-s), \ldots, (0,-1), (0,0), (0,1), \ldots , (0,s)\}$;

being s a small positive integer which determine the size of the structure element.



(a) (b)

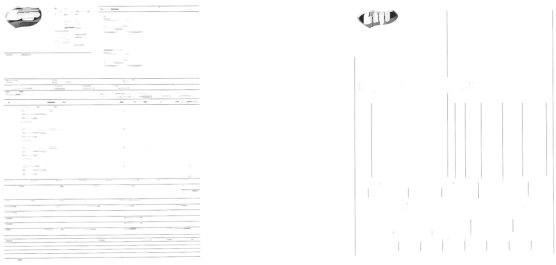
Figure 2. Structure elements (s=3): (a) B_{hor} , (b) B_{ver}.

In the preprocessing phase from the image I(x,y) two filtered images $I_{hor}=I_{hor}(x,y)$ and $I_{ver}=I_{ver}(x,y)$, which contains respectively horizontal and vertical segments, are obtained by a closure operator as follows (see Figure 3):

$$I_{hor} = I \bullet B_{hor} = (I \oplus B_{hor}) \Theta B_{hor} \quad (1a)$$

$$I_{ver} = I \bullet B_{ver} = (I \oplus B_{ver}) \Theta B_{ver} \quad (1b)$$

being "•" the closure operator, while "⊕" and "Θ" indicate respectively Minkowski sum and difference.



(a) (b)

Figure 3. Example of filtered images: (a) I_{hor} , (b) I_{ver}.

Finally, $I_{hor}(x,y)$ and $I_{ver}(x,y)$ are combined to reconstruct the preprocessed image I^* according to XOR operator:

$$I^* = I_{hor} \text{ XOR } I_{ver} \quad (2)$$

Figure 4 shows an example of document image after preprocessing.

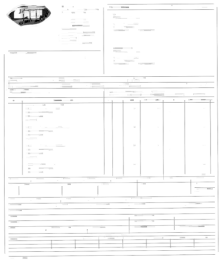


Figure 4. The preprocessed image I^*.

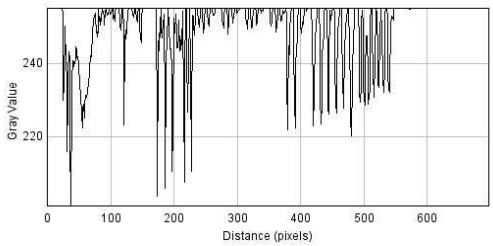
In the feature extraction step, in order to extract grid-based layout document images, the Radon Transform was considered. It is worth noting that the Radon Transform was extensively used in image analysis and has a number of important applications, like those related to MRI and computed tomography [Cormack, 1983; Deans, 1983]. The complete description of the Radon Transform is beyond the scope of this paper (see further details in [Jafari-Khouzani and Soltanian-Zadeh, 2005; Seo et al., 2004]). For the aim of this paper

we only remind that the Radon Transform computes projection sum of the image intensity along a oriented at line (ρ – xcos ϑ - ysin ϑ) =0, for each ϑ and ρ. More precisely the Radon Transform of a function $I^*(x,y)$ in an Euclidean space is defined by [Hjouj and Kammler, 2008]:

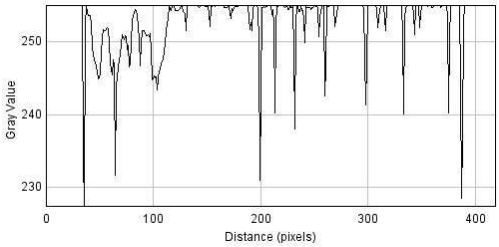
$$S_{\vartheta,\rho} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I^*(x,y)\cdot\delta(\rho - x\cos\vartheta - y\sin\vartheta)dxdy \quad (3)$$

where the d (r) is Dirac function, which is infinite for argument zero and zero for all other arguments (it integrates to one).

Therefore, computing the Radon Transform of a two dimensional image intensity function $I^*(x,y)$ results in its projections across the image at arbitrary orientations ϑ and offsets ρ. Figure 5 presents the results of the Radon Transform applied to the preprocessed image I^* for the parameter values related to horizontal (ϑ=0 , ρ=0) and vertical (ϑ=π/2 , ρ=0) projections.



(a) Horizontal Projection



(b) Vertical Projection

Figure 5. Feature extraction by Radon Transform.

Dynamic Time Warping (DTW) is used for matching the feature vectors extracted by the radon transform from two document images. More precisely, let be F^r, S^t the feature vectors of M elements extracted from the document images I^{*r} and I^{*t}, a warping function between S^r and S^t is any sequence of couples of indexes identifying points of S^r and S^t to be joined [Salvador and Chan, 2004; Lemire, 2009]:

$$W(S^r,S^t)=c_1,c_2,\ldots,c_K, \quad (4)$$

where $c_k=(i_k,j_k)$ $(k,i_k,j_k$ integers, $1\le k\le K$, $1\le i_k\le M$, $1\le j_k\le M)$. Now, if we consider a distance measure $d(c_k)=d(z^r(i_k), z^t(i_k))$ between elements of S^r and S^t, we can associate to $W(S^r,S^t)$ the dissimilarity measure

$$D_{W(S^r,S^t)} = \sum_{k=1}^{K} d(c_k) \quad (5)$$

The DTW detects the warping function $W^*(S^r,S^t)= c^*_1,c^*_2,\ldots,c^*_{K^*}$ which satisfies the condition of [Salvador and Chan, 2004]:

- Monotonicity (i.e. $i_{k-1} \leq i_k$, $j_{k-1} \leq j_k$ for k=2,...K) (6a)
- Continuity (i.e. $i_k - i_{k-1} \leq 1$, $j_k - j_{k-1} \leq 1$ for k=2,...K) (6b)
- Boundary (i.e. $i_1 = 1$, $j_1 = 1$ and $i_K = M$, $j_K = M$) (6c)

and which provides the distance value between S^r and S^t defined as [Salvador and Chan, 2004; Lemire, 2009]:

$$D_{W^*(S^r,S^t)} = \min_{W(S^r,S^t)} D_{W(S^r,S^t)} \quad (7)$$

.

The value in eq. (7) represents the similarity between the document images I^{*r} and I^{*t}. Therefore, given a document image as input, the matching module will outputs the ranked list of the k top similar document images retrieved from the database.

The matching procedure provides two distance-based ranked lists of documents, obtained respectively from horizontal and vertical projections. The decision making process obtains the final decision combining the two ranked lists using the Borda-count strategy [Kittler et al., 1998; Xu et al., 1992]. According to this strategy, let $D=\{D_1, D_2,\ldots, D_k,\ldots, D_K\}$ be the set of K documents enrolled into the system for reference and D" the unknown input document. Furthermore, let be:

- $L^h : < D^h_1, D^h_2, \ldots, D^h_k, \ldots, D^h_K >$ the ranked list of documents obtained from the match of the horizontal projection ($D^h_k \in D$, for k=1,2,,,,,K and $D^h_{k1} \neq D^h_{k2}$ for $k_1 \neq k_2$);
- $L^v: < D^v_1, D^v_2, \ldots, D^v_k, \ldots, D^v_K >$ the ranked list of documents obtained from the match of the vertical projection ($D^v_k \in D$, for k=1,2,,,,,K and $D^v_{k1} \neq D^v_{k2}$ for $k_1 \neq k_2$);

The Borda-count approach assigns to each reference document D_k a confidence score $S(D_k)$ defined as [Ho et al., 1994]:

$$S(D_k) = S^h(D_k) + S^v(D_k) \quad (8)$$

being $S^h(D_k)=K-i$, if $D_{k=} D^h_{i\ ;}$ $S^v(D_k)=K-j$, if $D_{k=} D^v_j$.

Hence, the final list of ranked documents is

$$L^* : < D^*_1, D^*_2, \ldots, D^*_k, \ldots, D^*_K > \quad (9)$$

so that D^*_{k1} precedes D^*_{k2} in L^* if and only is $S(D_{k1}) \geq S(D_{k2})$, and – of course - $_1$ is the top candidate document [Ho et al., 1994].

3. Experimental Results

The experimental test was carried out on a dataset of 33 commercial forms belonging to 16 different categories. Figure 6 shows some examples of commercial forms in the dataset. In this case they belong to the category n. 3.

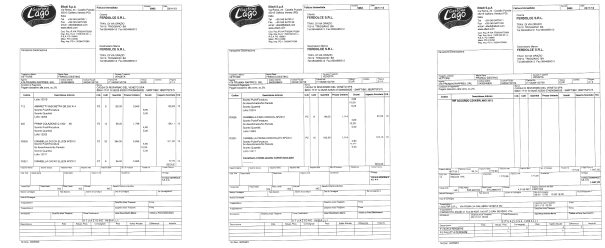


Figure 6. Examples of commercial forms of the same category.

Documents were scanned (100dpi , 256 gray-level) and preprocessed. Finally they were stored into a database along with the values of the Radon Transform concerning the horizontal ($S_{0,0}$) and vertical ($S_{0,\pi/2}$) projection. Table 1 reports the number of forms for each category.

Table 1. Dataset

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Number of Documents	9	4	3	3	2	2	1	1	1	1	1	1	1	1	1	1

In the testing phase the leave-one-out method was considered to verify the effectiveness of the system. In order to estimate the quality of the ranked list provided by the system for a given query, the Average Normalized Rank (ANR) was adopted, defined as follows [Huang et al., 2005]:

$$ANR = \frac{1}{N \cdot N_w} \cdot \sum_{i=1}^{N_w} \left(R_i - \frac{N_w + 1}{2} \right) \quad (10)$$

being

- N the number of documents in the set,
- N_w the number of relevant documents (for the given query) in the set,
- R_i is the rank of each relevant document in the set.

It is worst noting that ANR ranges in [0,1]:

- ANR=0 means that relevant documents are at the top of the ranked list (right position);
- ANR=1 means that relevant documents are at the bottom the ranked list (wrong position).

Figure 7 shows the experimental results. They demonstrate that the proposed approach is very robust with respect to different categories of documents. On average the value of ANR is equal to 0.08. Furthermore, 26 cases out of 33 the ANR is less than 10%, whereas only in one case it is greater than 0.5.

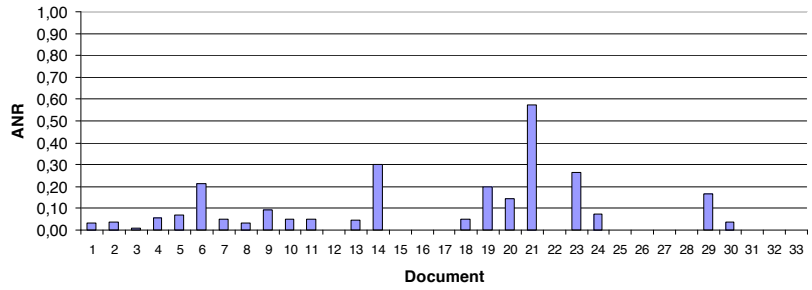


Figure 7. Experimental Results.

In order to estimate the robustness of the new approach two additional tests have been carried out using shifted and rotated document images as input.

When shifted documents are fed into the system the experimental results are shown in Figure 8. In this case a shift of 5 pixel is considered in the four main directions and the average result is computed. Also in this case the value of ANR is equal to 0.08, on average.

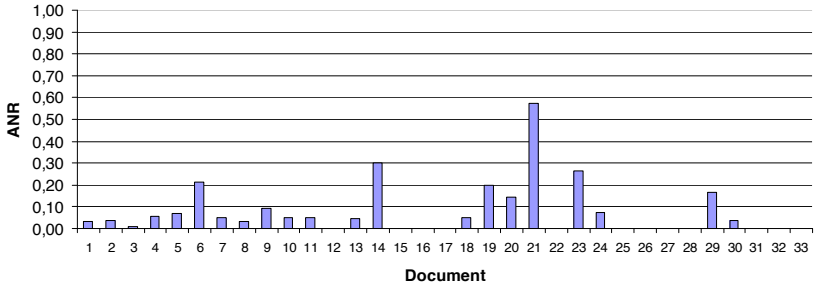


Figure 8. Experimental Results: shifted documents

Conversely, Figure 9 shows the results when rotated document images are fed into the system. In this case a rotation of 2° clockwise and anticlockwise is considered and the average result is computed. In this case the value of ANR is equal to 0.14 on average.

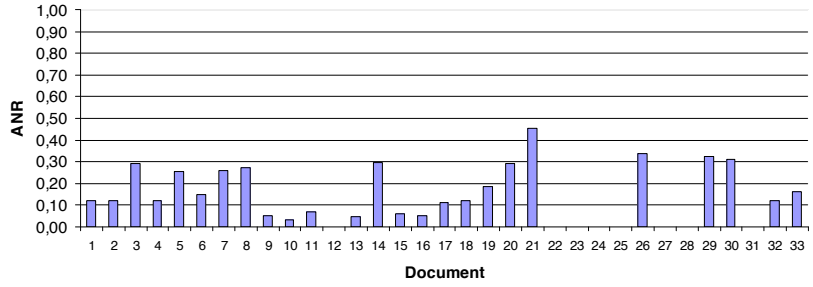


Figure 9. Experimental Results: rotated documents

Figure 10 shows the results when document images are shifted and rotated before to be fed into the system. In this case a shift of 5 pixels and a rotation of 2° clockwise and anticlockwise are considered. In t his case the value of ANR is equal to 0.15 on average.

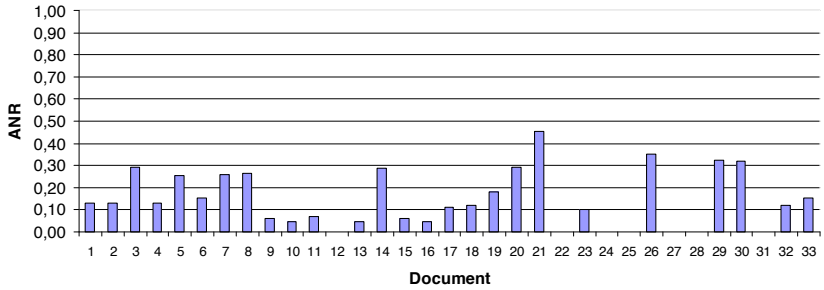


Figure 10. Experimental Results: shifted and rotated documents

4. Experimental Results

This paper presented a new system for layout-based document image retrieval. The system was specifically designed for retrieval of commercial forms as invoices, waybills and receipts, to optimize document management and sustainability. It used a morphologic filtering technique and the Radon Transform to obtain multiple document image descriptions. Document matching was then performed by Dynamic Time Warping whereas the Borda-count decision combination strategy was used to combine multiple decisions.

The experimental results, carried out on a dataset of real commercial documents, demonstrate the effectiveness of the proposed solutions and the robustness also with respect to shifted and rotated input document images.

5. References

Lopresti, D., Robust Retrieval of noisy text, *Proc. of ADL'96,* 1996, 76–85.

Cormack, A. M., Computed tomography: Some history and recent developments, *Proc. Symposia in Applied Mathematics*, 27, 1983, 35–42.

Deans, S. R., *The Radon Transform and Some of Its Applications*, New York: Wiley, 1983.

Doermann, D. , The Indexing and Retrieval of Document Images: A Survey, *Computer Vision and Image Understanding,* 70, 3, 1998, 287–298.

Duygulu, P., Atalay, V., A Hierarchical Representation of Form Documents for Identification and Retrieval, *International Journal on Document Analysis and Recognition,* 5, 1, 2002, 17–27.

Erol, B., Ant´unez, E., Hull, J. J., Hotpaper: multimedia interaction with paper using mobile phones, *Proceeding of the 16th ACM international conference on Multimedia*, 2008, 399–408.

Hjouj, F., Kammler, D. W., Identification of Reflected, Scaled, Translated, and Rotated Objects From Their Radon Projections, IEEE Trans. Image Processing, 17, 3, 2008, 301-310.

Ho, T.K., Hull, J.J., Srihari, S.N., Decision combination in multiple classifier systems, IEEE Trans. Pattern Anal. Mach. Intell., 16, 1, 1994, 66–75.

Huang, M., Dementhon, D., Doermann, D., Golebiowski, L., Document ranking by layout relevance, *Proc. 8th ICDAR*, 2005, 362–366.

Jafari-Khouzani, K., Soltanian-Zadeh, H., Radon Transform orientation estimation for rotation invariant texture analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, 27, 6, 2005, 1004–1008.

Kittler, J., Hatef, M., Duin, R.P.W., Matias, J., On combining classifiers, IEEE Trans. on Pattern Analysis Machine Intelligence, 20, 3, 1998, 226-239.

Lemire, D., Faster Retrieval with a Two-Pass Dynamic-Time-Warping Lower Bound, Pattern Recognition, 42, 9, 2009, 2169-2180.

Liu, Q., Liao, C., PaperUI, Proceeding of the 4th International Workshop on Camera-Based Document Analysis and Recognition, 2011, 3–10.

Manning, C. D. , Raghavan, P. , Schütze, H., An Introduction to Information Retrieval, Cambridge Press, 2009.

Marukawa, K., Hu, T., Fujisawa, H., Shima, Y., Document retrieval tolerating character recognition errors - Evaluation and application, *Pattern Recognition,* 30, 8, 1997, 1361-1371.

Mitra, M., Chaudhuri, B., Information retrieval from documents: A Survey, *Information Retrieval,* 2, 2/3, 2000, 141–163.

Salvador, S., Chan, P., Fast DTW: Toward Accurate Dynamic Time Warping in Linear Time and Space, *Proc. KDD Workshop on Mining Temporal and Sequential Data*, 2004, 70-80.

Seo, S., Haitsma, J., Kalker, T., Yoo, C. D., A robust image fingerprinting system using the Radon transforms, *Signal Process.: Image Commun.*, 19, 4, 2004, 325–339.

Serra, J., *Image Analysis and Mathematical Morphology*, Academic Press, 1982.

Taghva, K., Borsack, J., Condit, A., "Evaluation of model-based retrieval effectiveness with OCR text", *ACM TOIS,* 14, 1, 1996, 64–93.

Tzacheva, A., El-Sonbaty, Y., El-Kwae, A., Document Image Matching Using a Maximal Grid Approach, *Proceedings of the SPIE Document Recognition and Retrieval IX,* 2002, 121-128.

Xu, L., Krzyzak, A., Suen, C. Y., Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition, IEEE Transaction on Systems, Man and Cybernetics, 22, 3, 1992, 418-435.

Acquisitions, what will happen to my business? Predicting the acquired firm status

Bice Della Piana, Filomena Ferrucci, Gerardina Flammia, Carmine Gravino
Dipartimento di Studi e Ricerche Aziendali – Management and Information Technology
Università degli Studi Salerno
Via Giovanni Paolo II, 132 - 84084 - Fisciano (SA)
{bdellapiana, fferucci, gravino}@unisa.it

Abstract. *Mergers and acquisitions are of great practical importance in strategic, monetary, and social terms. They have attracted the research attention of several management disciplines. However, no studies have used prediction techniques for analyzing these challenging and high risk activities, particularly if we consider the viewpoint of the acquired firm. Guided by the firms' needs on a support to their strategic decisions, we analyzed the stability of the acquired firm's status. With the aim of predicting the evolution of these cooperative arrangements, we carried out an empirical study meant to verify the ability of some Machine Learning techniques to predict the acquired firm's status after two years from the acquisitions. In our analysis we employed 107 firms that were involved in acquisition, and applied a 10-fold cross validation. The preliminary results suggest that two of the employed techniques can be profitably exploited to provide accurate estimates of the acquired firm's status.*

Keywords: Acquisition, Machine Learning, Empirical Study

1.Introduction

Inter-firm cooperation is today a crucial part of management (Contractor & Lorange, 2002). The most common types of cooperative arrangements are short-term contracts, relational contracting, licensing, logistical supply-chain relationships, equity joint ventures and the complete merger or acquisition of two or more organizations (Gulati & Singh, 1998). Among these, mergers and acquisitions (M&A) are characterized by a high mutual commitment between partners, a high expected longevity of the cooperation and high risks. For these reasons, M&A are popular and controversial (Hayward, 2002) and have attracted the interest of a broad range of management disciplines encompassing the financial, strategic, behavioral, operational and cross-cultural aspects (Cartwright & Schoenberg, 2006) (Gomes et al., 2013).

The stability of the acquired firm's status in the post-acquisition stage (intended here as the status of the acquired firm that does not change over time) represents an important factor of this arrangement with relevant managerial implication. Considering the firms' needs on a support to their strategic decisions and the complexity of the multifaceted phenomenon of the acquisitions, it is useful to verify the possibility to predict the status of the firm acquired in the post acquisition stage. To this aim, we carried out an empirical study aiming to assess the prediction capability of some Machine Learning (ML) techniques. The application of ML techniques is not new in management sciences (for a review see Hakimpoor et al. 2011); indeed several problems have been addressed in different subject areas. Nevertheless, at the best of our knowledge, the prediction of the acquired firm status has not been investigated.

For the empirical analysis we employed as data set information about 107 acquisitions and analyzed the use of a subset of the classification techniques available in Weka (Hall et al., 2009). As for the validation method, we employed a 10-fold cross validation, while Precision, Recall, F-measures, Accuracy, and K-statistic were exploited as evaluation criteria.

The paper is organized as follows. Section 1 introduces the main research problem upon which this study is based and justifies the research methodology adopted. Section 2 describes the sources of data; Section 3 illustrates the design of the empirical analysis. Section 4 summarizes the findings of our investigation and highlights the future directions of the research.

2. Data set

The choice of the starting point for the identification of the firms being analyzed has not been easy to determine. There are many researches that utilize large alliance databases to explore issues related to firm collaboration (e.g., Folta and Miller, 2002; Hagedoorn, 2002; Lavie and Rosenkopf 2006; Vanhaverbeke, Duysters, and Noorderhaven, 2002; Villalonga and McGahan, 2005). Securities Data Company (SDC), MERIT-CATI, CORE, Bioscan, or Recombinant Capital (RECAP) are the alliance databases most commonly used in empirical studies published in top strategy journals (Shilling, 2009). Considering that our focus was the Italian context, these alliance databases were not useful for the purposes of the research. On the other hand the KPMG Report 2012 is very useful to understand the phenomenon of acquisitions in Italy since it aims to offer an across-the-board investigation of development trends in Italian business, by observing consolidation, internationalization, and competitive processes from a global standpoint. Nevertheless, this report was not helpful to our research goal since it does not provide the following information for both the acquired and the acquiring firms (name of the variables among brackets):

- general information about the firms: status as active or inactive (FirmsAcquiredStatus and FirmsAcquiringStatus, respectively), year in which the activity of the firms started, location of the headquarter (HLA1 and HLA2, respectively), legal capital (LegalcapitalofAcquired and LegalcapitalofAcquiring, respectively), value of production (ValueofProductionAcquired and

ValueofProductionAcquiring, respectively), number of employees (NumberofEmployeesAcquired and NumberofEmployeesAcquiring, respectively);

- specific information regarding the acquisitions: year of acquisition, specific constraints between firms.

As there is no database with similar features, it has been created ad hoc supported by the basic idea that it was necessary to have a data set with a certain degree of homogeneity among firms. To this aim, the most correct solution seemed to be to search for all Italian acquired firms which share the characteristic of being subject to the direction and coordination of other businesses (acquiring firms). In practice, the latter coordinates the strategic and operational decisions of the acquired firms. According to the art. 2497-bis of the Italian Civil Code, it has been specifically asked to select all the Italian companies that declared to the Chambers of Commerce to be involved in that type of acquisitions. Thus, the degree of homogeneity among firms is expressed by the communication of acquired firms of their situation of subjection to the management of the acquiring firms.

The research partnership between the Department of Management & Information Technology (University of Salerno) and the Chamber of Commerce allowed us to obtain information on the acquisitions carried out in Italy up to 2011. We obtained information on 26881 acquisitions in terms of the attributes/variables described above. We performed a preliminary analysis to remove those entries having null fields, obtaining as result 5.562 acquisitions. Moreover, before to process the data to address our research question, we carried out a data analysis with the aim of obtaining a more homogeneous sample and new attributes/variables needed for our prediction purposes. In particular, we introduced two further variables: Pre-acquisition Years of the Acquiring firm (PYA1) and Pre-acquisition Years of the Acquired firm (PYA2), where the PYA1 and PYA2 were obtained by taking into account the differences between the year in which the activity of the firms started and the year of acquisition (i.e., of being subject to the direction and coordination to other businesses - acquiring firms) communicated to the Chambers of Commerce.

From the selected 5562 acquisitions we could identify 4967 acquired firms having the status set to active, while the remaining 595 were acquired firms having status set to inactive. On the other hand, the analysis of the 5562 acquisitions allowed us to identify 5183 acquiring firms having status set to active, while the remaining 379 have status set to inactive. Moreover, we observed that more than 400 firms were characterized by PYA1 e PYA2 less than 7. The main part of the acquired firms was from regions in the north of Italy (i.e., Lombardia 1461, Veneto 774, Emilia Romagna 722) as well as the acquiring firms (i.e., Lombardia 1672, Veneto 753 Emilia Romagna 703), thus revealing that acquisitions are mainly generated among firms locating in the same region or in neighboring regions.

We also exploited the Principal Component Analysis (PCA) to perform a pre-analysis on the data to understand the relation among variables/attributes. PCA is widely used to reveal the internal structure of the data in a way that best explains the variance in the data (Abdi and Williams, 2010). The PCA results

showed that the variables positively correlated were: FirmsAcquiredStatus, FirmAcquiringStatus, LegalCapitalofAcquired, LegalCapitalofAcquiring, ValueofProductionAcquired, PYA1, PYA2, and NumberofEmployeesAcquired. On the other hand, FirmsAcquiredStatus and FirmAcquiringStatus were negatively correlated with NumberofEmployeesAcquiring and ValueofProductionAcquiring. Due to these results, we decided to exploit all these variables to address our research question. FirmsAcquiredStatus was the dependent variable, while the remaining variables were considered as independent.

Since we were interested in predicting the status of the acquired firms after two years from the acquisitions we needed the information on firms' status up to 2013 for all the firms obtained from the Chamber of Commerce (i.e., 5562 after preliminary analysis). This information can be achieved by manually querying the database made available by the Chamber of Commerce. Thus, we decided to randomly (according to the IP identification number ID) extract a number of firms (around 100) involved in acquisitions trying to obtain a balanced data set taking into account the information on the status. The result of our selection was: 57 firms tagged with an active status and 50 firms as inactive. Thus, we employed 107 observations (i.e., firms) in the empirical analysis, whose design is reported in the next section.

3. Design of the empirical analysis

We investigated some ML techniques to assess if it is possible to predict the status (active/inactive) of the firm involved in the acquisition as acquired. To this end, we formulated the following research question:

RQ: *Are the considered ML techniques able to predict the stability of an acquisition by estimating the status of the acquired firm?*

In the rest of this section we present the estimation techniques and the validation and evaluation criteria we employed in our empirical analysis. The results and their discussion are presented in Section 4.

3.1 Classification techniques

In this preliminary analysis we applied a subset of classification techniques available in Weka (Hall et al., 2009), which is a software "workbench" incorporating several classification techniques. It is worth noting that for all the employed techniques, no parameter setting was performed and we executed the default implementation provided by Weka (see Table 1). In the following we provide a brief description of these techniques. The reader interested to further details can refer to (Weka, 2013) and the references provided for the techniques.

Zero rule (**ZeroR).** It is the simplest of classifiers proposed by Weka, which considers no rules in its analysis and always outputs the most commonly found value for the dependent variable. For example, it predicts the mean if the type of the target attribute is numeric or the mode for an attribute whose type is nominal (Weka, 2013). It can be used as a lower bound on performance, i.e., a sort of benchmark, to assess the effectiveness of a proposed estimation technique.

Naive Bayesian Classifier (NBF). It is a simple but important probabilistic model (Caruana and Niculescu-Mizil, 2006), that uses Bayes's Theorem, i.e., assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It can be considered very efficient in a supervised learning setting. In many practical applications, parameter estimation uses the method of maximum likelihood.

Support Vector Machines (SVMs). SVMs are a maximum-margin linear classification technique (Vapnik, 1995), since they find the optimal hyperplane between two classes defined by a number of support vectors (Vapnik, 1995). This technique has good generalization ability, thanks to the introduction of a penalty factor, named C, which allows us to cope with the effects of outliers, permitting a certain amount of misclassification errors. Moreover, the use of a kernel function to map the data points into a higher-dimensional feature space allows SVMs to handle also non-linear problems (Vapnik, 1995). The parameter *epsilon* determines the level of accuracy of the approximation. Its value relies on the target values in the training set and the parameter C determines the trade-off between the occurrences of errors larger than epsilon in the training set and the flatness of the function. We chose to apply the Polynomial and the RBF kernels since they are widely used for prediction problems in different contexts (Scholkopf et al., 1997, Corazza et al., 2011, 2013).

MultiLayer Perceptrons (MLP). MLP is a feedforward artificial neural network that is trained with the backpropagation learning algorithm, a standard algorithm for any supervised-learning pattern recognition process (Miller et al., 2011). MLP uses three or more layers of neurons (nodes) with nonlinear activation functions and each layer fully connected to the next one. In our work we employed three layers, where the input layer is connected with the input data; the intermediate layer o hidden layer receives data from the input layer and sends them to the layer of output neurons; the last level or output layer receives data from neurons intermediate layer, and it interfaces with the output. As for the parameter settings (see Table 1), *HiddenLayers* represents the number of hidden neurons and there are some wildcards for this value as *a=(attribute+classes)/2*. *LearningRate* and *Momentum* are hyper parameters that determine the behavior of the algorithm, while *seed* is used for the selection of the starting point and for the shuffling of data. *TrainingTime* and *ValidationSetSize* represent the maximum number of iterations to execute and the number of patterns for training set that will not be used to find the weights. *ValidationThreshold* is a stopping criterion.

j48. It is the implementation of decision tree C4.5, a method of supervised classification structured in non-terminal and terminal nodes. The terminal nodes represent the results of the decision, while the non-terminal nodes indicate the test on one or more attributes (Zhao and Zhang, 2011). The algorithm tries to recursively partition the data set into subsets by evaluating the normalized information gain (difference in entropy) resulting from choosing a descriptor for splitting the data. The training process stops when the resulting nodes contain instances of single classes or if no descriptor can be found that would result to the information gain. *ConfidenceFactor* is used for pruning while *NumberObjects* allows setting the minimum number of instances before to fix a

leaf in the tree (set as reported in Table 1). *Seed* is used for randomizing the data when reduced error pruning is employed.

Simple Classification And Regression Trees (SCART). SCART is an algorithm to construct decision trees using historical data (Breiman et al., 1984). Through a pruning process, SCART uses cross-validation to split the input data into the two subgroups that are the most different with respect to the outcome. As for parameter setting (see Table 1), *MinNumObjects* represents the minimal number of observations at the terminal nodes while *NumFoldsPruning* is the number of folds in the internal cross-validation. *Seed* is the random number seed to be used and *SizePer* denotes the percentage of the training set size (in the range [0-1], where 0 means not included).

Table 1. The parameters employed for classifying

C	epsilon	Kernel,
1.0	1.0E-12	RBF with gamma=0.01
1.0	1.0E-12	Polynomial with c0=250007 and Exponent=1.0

(a) SVM

Hidden Layers	Random Seed	Training Time	Learning Rate	Momentum	Validation SetSize	Validation Threshold
a	16	1000	0.3	0.2	20	20

(b) MultiLayer Perceptrons

Confidence Factor	Number Objects	Seed
0.25	2	16

(c) J48

MinNumObjects	Random Seed	NumFoldsPruning	SizePer
2	16	10	1.0

(d) SCART

3.2 Validation method and evaluation criteria

We carried out a validation to verify whether or not ML techniques can provide useful predictions of the status of the acquired firm. To this end, we applied a 10-fold cross validation, randomly partitioning the original data set into 10 equal size subsets where a single subset is used as the validation set for assessing the predictions, and the remaining 9 subsets are used as training set to obtain the predictions. Thus, the validation process is repeated 10 times to use each of the 10 subsets exactly once as the validation set.

Regarding the evaluation criteria, we used several accuracy measures to assess the achieved predictions. In particular, we employed four performance measures, i.e., Accuracy, Precision, Recall, and F-measure (Witten I., Frank, 2005), which leverage on the concepts reported in the confusion matrix of Table 2 and are defined as follows. *Accuracy* is the ratio between the number of components correctly predicted (i.e., classified as TP and TN) and the total number of components (i.e., the sum of TP, TN, FP, FN). *Precision* is the ratio between the number of components classified as TP and the number of those classified as TP or FP. *Recall* is the ratio between the number of components classified as TP and the number of those classified as TP or FN

Table 2. The confusion matrix

		Predicted	
		Inactive	**Active**
Actual	**Inactive**	True Positive (TP)	False Negative (FN)
	Active	False Positive (TP)	True Negative (TN)

These definitions suggest that Precision concerns with the correctness of the responses provided by a method, while Recall allows us to measure the completeness of the responses. A measure that provides an indication of a balance between correctness and completeness is the harmonic mean of Precision and Recall (or *F-measure*), defined as follow:

$$F - measure = 2 * \frac{(\Pr ecision * \Re call)}{(\Pr ecision + \Re call)}$$

Moreover, we employed the K-statistic which is a chance-corrected measure of agreement between the classifications and the true classes (Weka, 2013). K-statistic is defined as:

$$K - statistic = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times the k raters agree and P(E) is the proportion of times the k raters are expected to agree by chance alone. K=1 means complete agreement while K=0 denotes lack of agreement (i.e. purely random coincidences of rates).

3.3 Validity evaluation

The validity of our empirical studies can be biased by different factors. First of all, threats to construct validity can be due to the way the variables have been selected and collected. We tried to mitigate such a threat employing publicly available data sets from the Chambers of Commerce. Threats to internal validity can be due to the bias introduced by the default choices considered in the application of the estimation techniques. However, our focus was not on the use of a specific technique but on the variables employed in combination with machine learning approaches to produce the predictions of the stability of an acquisition after the formation stage, by detecting the status (active/inactive) of the acquired firm. As for the external validity, it can be affected by the specific data set employed for the empirical analysis. However, this threat is mitigated by the fact that we randomly selected firms from the original data set provided by the Chambers of Commerce.

4. Results and discussion

Table 3 summarizes the results achieved in our empirical analysis in terms of K-statistics, Accuracy, Precision, Recall, and F-measure. We can observe that the decision tree based algorithms (i.e., NBF and SCART) provided predictions with the highest values for the considered performance measures

and with a high K-statistic value. In particular, the predictions on the status of the acquired firm in an acquisition with a Precision of about 0.87, thus with an incorrect estimates of about 13%. The predictions are also good in terms of completeness since the Recall value is 0.86 for NBF and SCART. Even if SVM_poly, MLP, and J48 achieved performance measure values close to the ones of NBF and SCART the K-statistic values suggest that this result was obtained by chance c.a. 50%. Thus, the performed analysis seems to reveal that the considered variables in combination with NBF and SCART are useful to predict the status of the acquired firms. More complex estimation techniques, such as SVMs, need accurate parameter settings "to give their best". Using the default setting of WEKA did not allow us to get useful predictions (with the RBF kernel the results are equal to the ones provided by ZeroR).

The results of this preliminary analysis are encouraging and motivate us to further investigate ML techniques in this context. We think that the error percentage characterizing the predictions can be reduced by optimizing the parameters setting.

Thus, we can positively answer our research question: *two of the employed ML techniques (i.e., NBF and SCART) are able to predict the stability of an acquisition by estimating the status of the acquired firm.*

Table 3. Results for each technique

Technique	K statistic	Performance Results			
		Accuracy	Precision	Recall	F-measure
ZeroR	0	0.533	0.284	0.533	0.370
NBF	**0.716**	**0.860**	**0.865**	**0.860**	**0.859**
SVM_Poly	0.436	0.729	0.820	0.729	0.701
SVM_RBF	0	0.533	0.284	0.533	0.370
MLP	0.51	0.757	0.757	0.757	0.756
J48	0.586	0.794	0.794	0.794	0.794
SCART	**0.716**	**0.860**	**0.865**	**0.860**	**0.859**

From a managerial perspective it is possible to highlight some of the strengths and weaknesses of the research. Referring to the objective of the research, knowing in advance the evolution of the firm acquired status could have relevant managerial implication. Strength of our work lies in the originality in terms of the specific condition (status) and specific phase (post acquisition) analyzed. None so far has bothered to understand what happens in the post-acquisition phase by analyzing the stability of the status of the acquired firm. The information about the stability of the status is particularly relevant. The prediction of the stability could be taken as an ex ante measure of the effectiveness of the strategy formulated by both the acquiring and acquired firms. Effectiveness intended as a measure of the chance to tie the actions, which in this case refers to the strategic choice of the acquisition, preferences and objectives of the companies involved. A weak point of our work involves the paucity of information on firms under investigation. It would be wise to have more information for both acquired and acquiring firms, as for example the change in sales and added value. Future developments will cover the use of prediction in order to analyze the coordination of activities between acquiring

and acquired firm whose status was predicted as active. As Agarwal et al. (2012) suggested, when one firm acquires another, there are numerous challenges in achieving coordination of activities between the acquired and acquiring entities.

As for future work, we intend to replicate the performed analysis by considering different and larger data sets and further estimation techniques. Furthermore, we are aware that the choice of parameters can have a strong impact on the application of the considered techniques, such as SVMs. In the presented preliminary study we applied the techniques with the default setting provided by Weka. In the future, we will search for the best parameter settings using approaches that allowed us to explore a wide range of variables' values.

References

[1] Abdi H., Williams L., Principal component analysis, in Wiley Interdisciplinary Reviews: Computational Statistics, 2, 4, July/August 2010, 433–459.

[2] Agarwal R, Croson R, Mahoney J. The role of incentives and communication in strategic alliances: an experimental investigation. Strategic Management Journal, 31(4), 2010, 413–437.

[3] Agarwal R., Anand J., Bercovitz J., Croson R., Spillovers across organizational architectures: the role of prior resource allocation and communication in post-acquisition coordination outcomes. Strategic Management Journal, 33, 2012, 710–733.

[4] Breiman L., Friedman J., Stone C., Olshen R., Classification and Regression Trees, 1984, Ed. Chapman and Hall/CRC.

[5] Cartwright S., Schoenberg R., Thirty years of mergers and acquisitions research: recent advances and future opportunities. British Journal of Management, 17(S1), 2006, S1-S5.

[6] Caruana R., Niculescu-Mizil A., An empirical comparison of supervised learning algorithms, Proceedings of the International Conference on Machine Learning, 2006.

[7] Conover W., Practical Nonparametric Statistics, 3rd edition. Wiley, 1998.

[8] Contractor F.J., Lorange P., The growth of alliances in the knowledge-based economy. International Business Review, 11, 2002, 485–502.

[9] Corazza A., Di Martino S., Ferrucci F., Gravino C., Mendes E., Investigating the use of support vector regression for web effort estimation. Empirical Software Engineering, 16, 2, 2011, 211–243.

[10] Corazza A., Di Martino S., Ferrucci F., Gravino C., Sarro F., Mendes E., Using Tabu Search to configure support vector regression for effort estimation. Empirical Software Engineering, 8, 3, 2013, 506-546.

[11] Folta T.B., Miller K.D. Real options in equity partnerships. Strategic Management Journal 23(1), 2002, 77–88.

[12] Gulati R, Nickerson J. Interorganizational trust, governance choice, and exchange performance. Organization Science, 19(5), 2008, 688–708.Sarkar M, Aulakh PS, Madhok A. Process capabilities and value generation in alliance portfolios. Organization Science, 20(3), 2009, 583–600.

[13] Gulati, R., & Singh, H. The architecture of cooperation: Managing coordination costs and Appropriation Concerns in Strategic Alliances. Administrative Science Quarterly, 1998, 43(4), 781–814.

[14] Hakimpoor H., Arshad K.A.B., Tat H.H., Khani N., Rahmandoust M., Artificial Neural Networks' application in management, World Applied Sciences Journal, 14(7), 2011, 1008-1019.

[15] Hall M., Eibe F., Holmes G., Pfahringer B., Reutemann P., Witten I., The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11, 1, 2009.

[16] Hagedoorn J., Inter-firm R&D partnerships: an overview of major trends and patterns since 1960. Research Policy, 2002, 31, 477–492.

[17] Hayward M.L.A., When do firms learn from their acquisition experience? Evidence from 1990-1995. Strategic Management Journal, 23, 2002, 21-39.

[18] Kitchenham B., Mendes E., Travassos G., Cross versus Within-Company Cost Estimation Studies: A Systematic Review. IEEE Trans. Software Eng. 33,5, 2007, 316-329.

[19] Lavie D., Rosenkopf L., Balancing exploration and exploitation in alliance formation. Academy of Management Journal 49, 2006, 797–818.

[20] Miller F., Vandome A., McBrewster J., Multilayer Perceptron: Artificial neural network, Activation function, Backpropagation, Linear separability, Perceptron, Supervised learning, Action potential, 2011, Ed.: Alphascript Publishing, ISBN-13: 978-6134149709

[21] Scholkopf B., Sung K., Burges C., Girosi F., Niyogi P., Poggio T., Vapnik V., Comparing support vector machines with Gaussian Kernels to radial basis function classifiers. IEEE Trans Signal Process 45, 11, 1997, 2758–2765.

[22] Vanhaverbeke W., Duysters G., Noorderhaven N., External technology sourcing through alliances or acquisitions: an analysis of the application-specific integrated circuits industry. Organization Science 13, 2002, 714–733.

[23] Vapnik V., The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[24] Weka Documentation, http://www.cs.waikato.ac.nz/ml/weka/documentation.html. Last check July 2013.

[25] Villalonga B., McGahan A.M., The choice among acquisitions, alliances, and divestitures. Strategic Management Journal 26(13), 2005, 1183–1208.

[26] Witten I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2nd Ed., San Francisco, 2005.

[27] Zhao Y., Zhang Y., Comparison of decision tree methods for finding active objects, Advances in Space Research.

Il ruolo delle start-up ICT-based nel cambiamento del paradigma di business: il caso delle smart grid[1]

Roberto Parente[1], Rosangela Feola[2], Valter Rassega[3]
[1]*Università di Salerno*
Via Giovanni Paolo II, 132 - Fisciano 84084 (SA)
rparente@unisa.it
[2]*Università di Salerno*
Via Giovanni Paolo II, 132 - Fisciano 84084 (SA)
rfeola@unisa.it
[3]*Regione Basilicata*
Ufficio Società dell'Informazione
Via V. Verrastro, 4 - 85100 Potenza (PZ)
walter@rassega.it

Abstract. *Business prospects related to the development of green technologies and energy saving systems have attracted the interest of ICT-based firms that have the crucial task of making more and more "intelligent" the system of production, distribution and use of the energy produced. In this process of radical redefinition of the energy business, an important role can be played by start-up that come from the world of research. The paper starts from a definition of the concept of smart grid as a trigger for radical change of the paradigm of the energy business. From a technological point of view we will examine in particular the role of ICT as enabling technologies in the development of smart grid. At the end we will analyze the role of ICT-based start-up in the development of smart grid, adding further evidence to the mainstream literature available on the role of new ventures in the development of radically new technologies.*

Keywords: Start-up; Technology Innovation; Smart grid

[1] Pur essendo il lavoro frutto della comune attività di ricerca degli Autori, il paragrafo 1 è attribuibile a Valter Rassega, il paragrafo 2 a Rosangela Feola, i paragrafi 3 e 4 a Roberto Parente. L'introduzione è attribuibile a tutti gli Autori. Gli Autori ringraziano la dott.ssa Valentina Cucino e il dott. Ezio Marinato per il supporto nella ricerca sugli spin-off italiani. Tuttavia, ogni responsabilità resta imputabile agli Autori.

Introduzione

Il paradigma tradizionale del business elettrico sta profondamente cambiando spostando l'attenzione dalle tradizionali fonti energetiche non rinnovabili verso fonti energetiche rinnovabili. Il sistema energetico è sempre più orientato al perseguimento di tre obiettivi: l'integrazione ottimale delle fonti rinnovabili nella rete di distribuzione elettrica, l'aumento dell'efficienza, della sicurezza e dell'affidabilità del servizio e la creazione di un mercato competitivo dell'energia. Il passaggio che ormai si impone è quello verso la **smart grid**, ovvero "una rete di trasporto e distribuzione di energia elettrica di nuova generazione in grado di integrare intelligentemente il comportamento e le azioni di tutti gli utenti ad essa connessi – produttori, consumatori e produttori/consumatori – allo scopo di assicurare un approvvigionamento elettrico efficiente, sostenibile, economico e sicuro" [European Technology Platform «Smartgrids», 2006]. Nello sviluppo delle smart grid, le tecnologie ICT rappresentano un elemento essenziale: le tecnologie informatiche rappresentano infatti una tecnologia abilitante in grado di conferire al sistema le caratteristiche di "intelligenza" necessarie a svolgere le funzioni "smart" della rete. Nel processo di ridefinizione del business dell'energia che si caratterizza per l'interrelazione che viene a crearsi tra il settore elettrico ed il settore ICT, diviene fondamentale indagare il ruolo che le imprese, vecchie e nuove, possono giocare nel processo di cambiamento del settore. In questo contesto, il paper si ripropone di mettere in luce il ruolo delle start-up ICT provenienti dal mondo della ricerca nello sviluppo delle smart grid. Dal punto di vista teorico il lavoro può essere inquadrato nell'ambito degli studi che analizzano i processi di trasformazione dei settori economici, ed in particolare di quelli che si sono focalizzati sull'analisi dei fattori e degli attori che guidano il cambiamento [Geels, 2002; Geels, 2004; Geels e Schot, 2007; Geels e Raven, 2006; Dolata, 2009; Erlinghagen e Markard, 2012].

1. Smart grid: verso un nuovo paradigma nel settore energetico

L'attuale evoluzione del mercato energetico si caratterizza per la regolazione delle strutture di mercato per l'energia elettrica che devono integrare agevolmente nella rete un numero crescente di produttori di energia elettrica, anche di piccola taglia, proveniente da fonti rinnovabili: vengono introdotti, in questo modo, nuovi concetti, come la gestione e modulazione della domanda di energia elettrica da parte dell'utilizzatore finale del servizio. Non a caso, una delle funzioni fondamentali della smart grid è quella di integrare nel sistema elettrico la micro generazione di energia elettrica operata da privati che, con i loro impianti fotovoltaici o eolici, potendo conferire in rete il loro surplus energetico, assumono la doppia veste di consumatore e di produttore: per l'appunto, i c.d. "prosumers". Per quanto limitata in termini di potenza prodotta per singolo impianto, la prevedibile diffusione su vasta scala di impianti di generazione di tipo casalingo segna l'avvento di un nuovo paradigma economico: il passaggio da un mercato di monopolio naturale integrato verticalmente, ad un mercato orizzontale dove potenzialmente tutti possono consumare e produrre senza limitazioni tecniche o normative. Quando produrre e vendere elettricità diventa più semplice, si attivano scambi e si crea uno

spazio di business per nuovi soggetti economici che offrono servizi di aggregazione. L'aggregatore è una delle figure emergenti nel nuovo panorama del settore energetico, e svolge la funzione di mediatore tra i consumatori, ai quali vende le flessibilità di carico ovvero la disponibilità a modulare i propri consumi energetici, e il mercato. In tal modo, l'aggregatore risponde alle richieste provenienti dal mercato e dal gestore del sistema di distribuzione, cercando di soddisfarle, utilizzando e aggregando la capacità di poter variare gli assorbimenti e, parallelamente, valutando le disponibilità energetiche fornite da un aggregato di prosumers. Dunque, l'attuale mercato dell'energia si configura come un ambito di interazione che vede un insieme composito di portatori di interesse. In chiave economica, deve essere analizzata tale interazione applicata al modello di tecnologia di rete della smart grid. Per gli economisti le tecnologie di rete sono quelle infrastrutture tecnologiche che connettono più agenti in rete coinvolti in un'interazione, e per le quali esistono delle esternalità positive. Detto in altri termini, ciò significa che il beneficio per il singolo agente non dipende solo dalla sua adozione individuale della tecnologia, ma è proporzionale al numero di altri agenti che entrano nel gioco. La smart grid sarà utile ai vari stakeholders solo se tutti i players di mercato parteciperanno alla loro implementazione, e l'utilità sarà tanto maggiore quanto più il modello diventerà diffuso e pervasivo. La transizione verso il nuovo paradigma sembra ormai avviata, ma sono subito emerse alcune barriere che è necessario superare. Una fra queste è quella delle infrastrutture di trasporto e delle connessioni di rete, che devono modificare le proprie caratteristiche e modalità operative per poter gestire flussi di energia multi-direzionali ed integrare in rete i sempre più numerosi impianti di generazione di energia elettrica. Per rispondere a queste esigenze, è necessario che i dispositivi di controllo e di gestione di tutti gli impianti di produzione e distribuzione elettrica, possano dialogare, sincronizzarsi e coordinarsi automaticamente fra loro per provvedere al trasporto e alla distribuzione dell'energia elettrica con la massima efficienza ed affidabilità. Una rete intelligente dunque fortemente caratterizzata dal ruolo fondamentale svolto dalle tecnologie ICT, grazie alle quali i dispositivi e gli apparati di controllo e gestione della rete elettrica, gli utilizzatori, i produttori e ogni altra figura coinvolta nel network possono scambiarsi informazioni in tempo reale, determinando lo sconvolgimento dei tradizionali paradigmi di settore e le conseguenti modificazioni degli scenari economici e tecnologici nel campo energetico. Non a caso, nella prospettiva economica, la smart grid può essere considerata un agente "transattivo" che permette di effettuare transazioni di natura finanziaria e informativa tra tutti gli attori del mercato energetico.

Per ritenersi "Smart", la grid di nuova generazione deve necessariamente avere alcune caratteristiche fondamentali: deve essere innanzitutto un diffuso e potente sistema di elaborazione computerizzato, connesso con i dispositivi di misura e le apparecchiature di controllo della rete, in grado di effettuare misure, calcoli e comunicare in tempo reale con gli altri sistemi di elaborazione costituenti l'arcipelago tecnologico della rete stessa. La tecnologia necessaria è già disponibile e alcuni degli apparati di gestione della grid sono già installati e in esercizio sulle reti di distribuzione ad alta e media tensione e molti produttori stanno rilasciando nuovi dispositivi elettronici di potenza per il controllo del

flusso energetico a livello di distribuzione a bassa tensione. Tuttavia, va rilevato che tali dispositivi, per concorrere realmente alla realizzazione di una rete intelligente, devono integrarsi perfettamente e, anche se molto è già stato fatto, ad oggi si devono ancora definire alcuni standard necessari a supportare un livello di interoperabilità hardware e software di tipo "plug-and-play" tra tutti i dispositivi elettronici utilizzabili nella smart grid. In termini operativi, la smart grid è in grado di conoscere e profilare le nostre abitudini nell'ambito dei consumi elettrici e il fabbisogno degli elettrodomestici e degli eventuali veicoli elettrici posseduti. Quando nella rete la richiesta di elettricità è al massimo, la Smart grid si adatta automaticamente alla domanda raccogliendo l'energia in eccesso disponibile da altre fonti connesse in rete, evitando problemi di sovraccarico. Oltretutto, il sistema elettrico Smart supporta e gestisce comunicazioni in tempo reale tra i consumatori e i gestori della distribuzione e ciò fornisce automaticamente agli utenti le informazioni e gli strumenti per il controllo e l'esercizio delle opzioni che rendono possibile adattare dinamicamente il proprio profilo di consumo di energia in funzione dell'andamento del prezzo o, nel caso di utenti che dispongono di un proprio impianto di generazione eolica o fotovoltaica, il sistema consente agli stessi di intervenire attivamente nel mercato elettrico immettendo in rete il proprio surplus di energia prodotta, ad un prezzo di vendita determinato dinamicamente dal sistema. Il percorso verso l'effettiva e completa realizzazione della smart grid appare solo agli inizi e la strada da compiere è ancora lunga. Il mercato delle apparecchiature e dei servizi per la costruzione di una rete sempre più intelligente è in forte crescita ed evidenzia un trend positivo che non è soltanto italiano ma che è diffuso anche a livello internazionale.

2. Il ruolo delle tecnologie ICT nello sviluppo delle smart grid

Nel passaggio dal vecchio al nuovo paradigma energetico il ruolo delle tecnologie ICT risulta sempre più determinante. Le tecnologie ICT rappresentano infatti il fattore abilitante di un processo di trasformazione strutturale di ogni fase del ciclo energetico, dalla generazione all'accumulo, al trasporto, alla distribuzione, alla vendita e, soprattutto, al consumo intelligente di energia. Le tecnologie ICT rappresentano l'elemento in grado di conferire al sistema le caratteristiche di "intelligenza" necessarie a svolgere le funzioni "smart" della rete e di poter operare il controllo automatico e adattivo delle infrastrutture in funzione delle condizioni operative del momento. Il connubio tra smart grid e tecnologie ICT è implicito nell'essenza stessa del processo di realizzazione della smart grid: realizzare una smart grid non significa infatti rifare da zero le infrastrutture della rete elettrica ma significa arricchire di intelligenza quelle esistenti. Si potrebbe parlare di un Internet of Energy in cui una rete di informazione si affianca ad una rete di distribuzione fisica dell'energia consentendone una gestione intelligente. Le tecnologie ICT intervengono nello sviluppo della smart grid a due livelli [Bellifemine et al, 2009]: forniscono gli strumenti, le piattaforme e gli algoritmi di controllo distribuito, necessari ad ottimizzare l'efficienza di tutti i sistemi coinvolti; divengono il fattore abilitante per lo sviluppo di una serie di servizi legati all'energia.

Nello sviluppo di una rete energetica intelligente, le tecnologie ICT consentono di rispondere a tre esigenze fondamentali: favorire l'integrazione

della generazione distribuita; facilitare la gestione dei flussi multi direzionali; favorire il coinvolgimento del cliente.

Va tuttavia precisato che molte delle tecnologie ICT necessarie per le smart grid sono oggi disponibili come elementi separati, sviluppati a diversi gradi di maturità. Si rendono necessari ulteriori investimenti in R&S finalizzati a sviluppare le tecnologie funzionali alla smart grid portandole a un livello di maturità adeguato a poterle installare su vasta scala. Il potenziale del mercato ICT collegato agli investimenti nelle reti intelligenti è estremamente elevato come dimostra il crescente coinvolgimento delle imprese ICT nei progetti europei relativi alle smart grid. Se si analizza infatti l'andamento della partecipazione delle imprese ICT ai progetti europei, si riscontra, dal 2004 ad oggi una tendenza crescente: da una percentuale di circa il 20% di progetti partecipati da almeno un'impresa ICT nel 2004, si arriva ad una percentuale di oltre il 60% nel 2012 [EU, 2013]. La consapevolezza del ruolo fondamentale delle tecnologie informatiche è testimoniata, sul fronte istituzionale, dall'avvio di numerose iniziative europee (quali il progetto SEESGEN ed il progetto Finsey) che hanno l'obiettivo di potenziare il ruolo dell'information technology e delle telecomunicazioni per migliorare l'efficienza energetica nella produzione e distribuzione di energia. Le prospettive di business legate alle applicazioni ICT nel campo delle smart grid hanno richiamato l'interesse dei grandi operatori del settore (Telecom, Cisco, Ibm, Microsoft e Google) che hanno effettuato notevoli investimenti nella direzione smart e nella direzione dello sviluppo di dispositivi e tecnologie informatiche necessarie alla implementazione di un rete energetica intelligente. Tuttavia, il processo è appena agli inizi e la strada da percorrere verso la completa implementazione di una rete energetica intelligente è ancora lunga: il progetto "European Smart Grids Technology Platform" della Commissione Europea stima che siano ancora necessari investimenti per 750 miliardi di Euro nei prossimi trent'anni, di cui 100 miliardi nella trasmissione, 300 miliardi nella distribuzione e 350 miliardi nella generazione.

Il tema del ruolo dell'ICT nello sviluppo delle smart grid, può essere inquadrato nell'ambito degli studi che analizzano i processi di trasformazione dei settori economici, ed in particolare di quelli che si sono focalizzati sull'analisi dei fattori e degli attori che guidano il cambiamento [Geels, 2002; Geels, 2004; Geels e Schot, 2007; Geels e Raven, 2006; Dolata, 2009]. A questo proposito è stato riconosciuto che un ruolo determinante nella trasformazione di un settore può essere giocato da imprese consolidate operanti in settori magari anche tecnologicamente distanti da quello di riferimento ma portatori di competenze e tecnologie in grado di innescare processi di cambiamento [Erlinghagen e Markard, 2012]. Tali incumbents infatti potrebbero decidere di entrare in un nuovo settore con l'obiettivo di sfruttare al meglio la già esistente base di conoscenze e competenze [Niccholls-Nixon e Jasinski, 1995]. Il ruolo di tali incumbents diventa particolarmente rilevante quando le tecnologie sviluppate all'esterno del settore di riferimento (quelle che potremmo definire come outside technologies) diventano abilitanti per la trasformazione del settore stesso [Dolata, 2009]. In questi casi infatti, settori precedentemente lontani e completamente separati, divengono strettamente interconnessi e dipendenti gli uni dagli altri [Erlinghagen e Markard, 2012]. L'interrelazione tra settori che

viene a determinarsi quando la forza del cambiamento è esterna al settore di riferimento, ha importanti conseguenze sulle dinamiche competitive del settore stesso. In questi contesti infatti, il gioco competitivo coinvolge non solo gli incumbents del settore ed i nuovi entranti, ma anche gli incumbents dei settori adiacenti, che potrebbero decidere di entrare nel focal sector per sfruttare le opportunità che in esso si presentano. L'ingresso degli incumbents adiacenti è in grado di alterare il "naturale gioco competitivo" che come dimostra la letteratura sull'argomento, generalmente vede gli incumbents del settore soccombere di fronte alla forza innovatrice dei new-comers [Hannan e Freeman, 1977]. Di fronte alla minaccia derivante dagli incumbents adiacenti, le imprese consolidate nel settore focale, piuttosto che resistere al cambiamento, tendono a reagire adottando strategie offensive con l'obiettivo di acquisire, ad esempio attraverso acquisizioni di nuove imprese, le competenze e le tecnologie che non possiedono al proprio interno [Christensen, 1997].

Il settore delle smart grid è perfettamente rappresentativo di tali dinamiche competitive. Lo sviluppo di una rete intelligente infatti, implica una profonda trasformazione del settore energetico derivante dall'applicazione al suo interno delle conoscenze e delle tecnologie derivanti dal settore ICT. Lo sviluppo delle smart grid dipenderà quindi dai comportamenti e dalle strategie che saranno adottate dagli incumbents elettrici, dagli operatori incumbents del settore ICT e da nuovi entranti sia nel settore elettrico che nel settore ICT.

La nostra analisi si focalizzerà in particolare sul ruolo che la ricerca e le start-up ICT possono svolgere nel cambiamento del paradigma energetico.

3. Il contributo della ricerca e delle start-up nello sviluppo del nuovo paradigma energetico

Shock esogeni al sistema produttivo e nuove acquisizioni scientifiche e tecnologiche rappresentano le driving forces dei cambiamenti radicali nella struttura dei settori industriali [Geels, 2004]. Ovviamente, pur essendo possibile che nuove scoperte scientifiche e nuove tecnologie nascano sulla scia di una ricerca curiosity driven, è ancor più verosimile che siano shock esogeni al sistema produttivo, com'è nel caso del cambiamento climatico in corso, ad agire quale elemento catalizzatore nello sviluppo di nuove tecnologie destinate a cambiare i connotati dei processi produttivi tradizionali [Schaltegger e Wagner, 2011]. La direzione di questo cambiamento, ed i tempi necessari per portarlo a compimento, dipendono da una molteplicità di attori e dai comportamenti che essi mettono in campo, dando vita ad una complessa rete di influenze e di condizionamenti reciproci. Le imprese rappresentano gli attori chiave del cambiamento, giacché in ultima analisi sono queste peculiari strutture organizzative che, alla ricerca delle più idonee condizioni di sviluppo e di profittabilità, sono chiamate a dover elaborare una sintesi fra bisogni del mercato e tecnologie disponibili, attraverso una loro peculiare offerta di prodotti/servizi. In un sistema competitivo aperto è pacifico che tipologie diverse di imprese possano avere atteggiamenti, capacità, valutazioni diverse (anche fortemente divergenti) sia rispetto alle strategie da adottare che ai tempi della loro implementazione [Denrell et al, 2003]. A questo proposito, un tema ben noto in letteratura è legato al diverso atteggiamento nei confronti delle

innovazioni radicali espresso dalle aziende già consolidate (i cosiddetti incumbents) rispetto alle aziende di nuova formazione (le cosiddette start-up). Per spiegare tale fenomeno si è fatto ricorso al concetto di nicchia ecologica e di inerzia organizzativa [Hannan e Freeman, 1977]. Le imprese, cioè, che hanno avuto successo in un determinato ambiente tecnologico difficilmente hanno la forza e la capacità di sostituire/modificare radicalmente le proprie competenze tecnologiche e finiscono così per subire un processo di "adverse selection" ovvero ad essere sostituite da coorti di nuove imprese che nascono fin dall'inizio cavalcando le nuove tecnologie [Hill e Rothaermel, 2003]. Dosi ed altri [2008], analizzando l'impatto delle nuove tecnologie ICT sulla struttura competitiva dei settori economici hanno verificato che la rivoluzione tecnologica delle ICT ha modificato solo marginalmente la struttura dei settori industriali. I confini verticali ed orizzontali delle imprese sono cambiati, c'è stato un certo turnover nel club delle grandi imprese, ma non si osserva una crisi delle grandi imprese a favore di quelle più piccole, e più specializzate, che verosimilmente coincidono in buona parte con la categoria delle imprese più giovani. Una delle spiegazioni plausibili di questa apparente contraddizione risiede nella capacità che hanno mostrato le grandi imprese che hanno abbracciato la filosofia dell'Open Innovation [Chesbrough, 2008], di saper integrare al loro interno asset tecnologici e competenze sviluppate al loro esterno, magari da giovani start-up. Un indizio in questo senso viene portato da Hockerts e Wustenhagen [2010], che, analizzando in particolare i cambiamenti in corso nel settore delle energie rinnovabili, hanno evidenziato che nell'ambito di queste nuove tecnologie svolgono un ruolo portante sia gli "Emerging Davids" (ovvero le start-up tecnologiche), sia i "Greening Goliath" (ovvero le vecchie imprese capaci di innovarsi). I rapporti fra nuove e vecchie imprese sono quindi più complessi e vanno oltre il paradigma sostituto/sostituito, con le nuove imprese technology-based che possono giocare un ruolo più indiretto di agenti di cambiamento e di ringiovanimento tecnologico del sistema produttivo preesistente [Autio e Yli-Renko, 1998]. Il ruolo di ringiovanimento dei Big Incumbents è documentato anche nell'ambito specifico delle tecnologie smart grid [Erlinghagen e Markard, 2012]. Questi autori evidenziano una triangolazione fra Grandi imprese del settore ICT, nuove start-up ICT e Grandi imprese delle vecchie Utilities. Ad un ingresso nelle tecnologie smart grid delle grandi imprese ICT, le Grandi utilities hanno infatti risposto accelerando i processi di acquisizione di nuove start-up specializzate nelle tecnologie ICT. Questo processo prosegue tuttora e trae linfa anche da nuovi e sempre più sofisticati sistemi di scouting dei progetti di nuove start-up tecnologiche. Incubatori, Acceleratori, Start cup Competition, Seed Capital sono soltanto alcuni dei filoni di questo nuovo attivismo dei Big Players per attrarre giovani start-up. Per fermarci al contesto italiano Enel Lab e Working Capital promosso da Telecom Italia sono alcuni degli esempi più interessanti in questo campo.

Quando si parla, però, di innovazioni radicali che nascono quindi sui confini della ricerca scientifica entra in gioco un ulteriore attore che diventa fondamentale, ovvero il sistema della Ricerca pubblica rappresentato dalle Università e dai Centri di ricerca pubblici. Il loro impegno nella ricerca scientifica sulle aree di frontiera e nel trasferimento tecnologico diventa determinante.

L'impegno di questi soggetti si può misurare attraverso una serie di indicatori, quali il numero di pubblicazioni scientifiche in un determinato ambito scientifico/tecnologico, ma soprattutto il numero di brevetti e di spin-off attivati a valle del lavoro svolto dai gruppi di ricerca. Per ciò che concerne le pubblicazioni scientifiche, il rapporto pubblicato dall'Istituto per la Competitività per il settore delle Smart Grid rileva per il 2012, 199 pubblicazioni a livello mondiale, a fronte di 86 pubblicazioni nel 2011 e di sole 18 pubblicazioni nel 2010 [I-Com, 2013]. L'Italia, nell'anno 2012, con le sue 8 pubblicazioni, si attesta nella media in questo campo, in forte crescita rispetto al dato del 2011 (solo 3 pubblicazioni). Da notare che all'interno delle smart grid, la ricerca italiana si è focalizzata sull'importante settore dello *Smart Metering:* l'Italia nel 2012 con 6 pubblicazioni (erano 2 nel 2012) si attesta come primo Paese seguita da Germania e Cina. Analizzando invece il numero di brevetti, l'Italia accusa un significativo ritardo. Nel 2012 risultano complessivamente 171 brevetti a fronte di 96 nel 2010, 34 nel 2009 e solo 10 brevetti nel 2000. L'Italia nel 2012 presenta solo 2 brevetti che sono di provenienza lombarda e registrati da aziende. Per quanto riguarda il ruolo degli spin-off accademici ICT-based, attraverso una ricerca on desk condotta principalmente mediante consultazione dei siti web di 967 spin-off universitari, sul totale di 1.082 censiti dal Netval nel 2012 [Netval, 2013] si è osservato in primo luogo che circa il 30% di tali spin-off hanno matrice di provenienza ICT. Su questo gruppo di spin-off si è proceduto ad una analisi più approfondita per verificare se operano in almeno una delle aree classificate dall'OECD come aree di applicazione delle tecnologie ICT al settore Smart Grid. La classificazione OECD distingue 5 macroaree: Generation, Transport, Storage, Retail, Consumption [OECD, 2011]. Sulla base di questa catalogazione sono state censite complessivamente 44 imprese spin-off con tecnologie e competenze applicabili ad almeno una di queste macroaree. All'interno di questo gruppo, 27 spin-off pur possedendo le potenzialità per operare nel settore smart grid al momento sono attive solo in settori affini. 17 imprese invece dichiarano di essere già operative in aree applicative delle tecnologie ICT alle smart grid. Le aree specifiche di attività dei 17 spin-off ICT già impegnati nelle smart grid sono riportate nella Fig. 2.

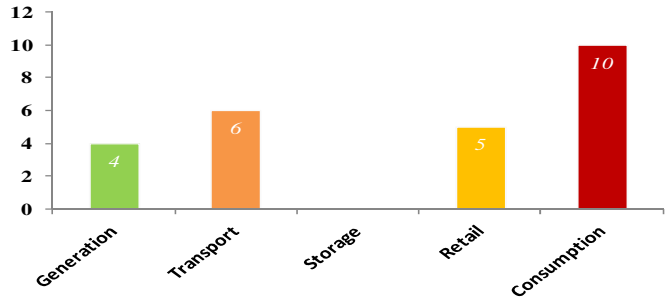


Fig. 2 – Aree di attività degli spin-off nelle smart grid

Come si evince dalla figura, l'area più presidiata è quella "Consumption", con 10 aziende presenti. Da rilevare che diversi spin-off operano contemporaneamente in più aree di business.

4. Conclusioni

Il mondo della ricerca è dunque fortemente coinvolto nella produzione di nuove conoscenze scientifiche e tecnologiche da applicare al settore delle smart grid, come testimonia il volume crescente di pubblicazioni in questo campo, ma sembrerebbe, almeno in Italia, ancora poco interessato al tema della valorizzazione economica considerata la scarsa propensione alla brevettazione. Se ci si fermasse a questi dati ne verrebbe però fuori un'immagine distorta e poco realistica. Molto spesso, infatti, la valorizzazione delle nuove conoscenze scientifiche e tecnologiche prodotte dai gruppi accademici viene perseguita direttamente tramite la costituzione di spin-off accademici che poi evidentemente nel corso della loro operatività procedono alla eventuale brevettazione del proprio know how. Il numero di spin-off ICT-based in Italia operanti nel settore smart grid è, infatti, significativo in termini assoluti e se comparati con il volume di pubblicazioni scientifiche e, soprattutto di brevetti Universitari prodotti. Il fiorire, anche in Italia, di spin-off che si propongono di applicare tecnologie ICT-based al settore smart grid si presta ad una chiave di lettura che fa riferimento alle difficoltà legate al trasferimento tecnologico diretto verso i cosiddetti "incumbents" [Cerrato et al, 2011]. Questa chiave di lettura mette in luce la difficoltà degli incumbents a gestire cambi di paradigmi tecnologici che sono "competence destroying" e che minano quindi alle fondamenta, gli assetti organizzativi e tecnologici che sono stati alla base del successo nel passato, e per converso la maggiore capacità delle new ventures ad adattarsi ai cambiamenti repentini che si verificano all'interno di rugged competitive landscapes [Levinthal, 1997] quale è, appunto, il caso delle smart grid technologies. Vi sono rilevanti indizi, però, per ritenere che le start-up ed in particolare gli spin-off accademici possano rappresentare per gli incumbents più "smart" delle opportunità per ampliare/modificare la propria base di competenze e conoscenze utili a dominare il nuovo paradigma tecnologico che sta emergendo con la smart grid.

Bibliografia

Autio, E., Yli-Renko, H., New technology-based as agents of technological rejuvenation. Entrepreneurship & Regional development, 10, 1998, 71-92.

Bellifemine, F. L., Borean, C., De Bonis, R., Smart grids: Energia e ICT. Notiziario Tecnico Telecom Italia, Anno 18, Num. 3, 2009.

Cerrato, D., Parente, R., Petrone, M., La collaborazione tra università e industria: un'indagine sui brevetti co-generati in Italia. L'Industria. Vol. A. XXXIII, N.2, 2012, 255-282.

Chesbrough H., Open. Modelli di business per l'innovazione, a cura di A. Di Minin, tradotto da R. Merlini, Egea, Milano, 2008.

Christensen C.M., The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail, Harvard Business School Press, Boston, MA, 1997.

Denrell, J., Fang, C., Winter, S., The economics of strategic opportunity. Strategic Management Journal, 24, 2003.

Dolata, U., Technological innovations and sectoral change: transformative capacity, adaptability, patterns of change: an analytical framework. Research Policy, 38, 2009, 1066–1076.

Dosi, G., Gambardella, A., Grazzi, M., Orsenigo, L., Technological revolutions and the evolution of industrial structures. Capitalism and society, 3, 2008.

Erlinghagen, S., Markard, J., Smart grid and the transformation of the electricity sector: ICT firms as potential catalyst for sectoral change. Energy Policy, 51, 2012, 895-906.

EU-European Commission, Smart Grid projects in Europe: Lessons learned and current developments, Report by the Joint Research Centre of European Commission, 2013.

European Technology Platform (ETP) Smartgrids, Vision and Strategy for Europe's Electricity Networks of the Future, Directorate-General for Research, Bruxelles, 2006.

Geels, F.W., From sectoral systems of innovation to socio-technical systems: insights about dynamics and change from sociology and institutional theory. Research Policy, 33, 2004, 897–920.

Geels, F.W., Raven, R., Non-linearity and expectations in niche-development trajectories: ups and downs in Dutch biogas development (1973–2003). Technology Analysis & Strategic Management, 18, 2006, 375–392.

Geels, F.W., Schot, J., Typology of socio technical transition pathways. Research Policy, 36, 2007, 399–417.

Geels, F.W., Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study, Research Policy, 31, 2002, 1257–1274.

Hannan, M., J. Freeman, The population ecology of organizations. American Journal of Sociology, 82, 5, 1977, 929-964.

Hill, C.W.L., Rothaermel F.T., The performance of incumbent firms in the face of radical technological innovation. Academy of Management Review, 28, 2, 2003, 257-274.

Hockerts, K., Wüstenhagen R., Greening Goliaths versus emerging David. Theorizing about the role of incumbents and new entrants in sustainable entrepreneurship. Journal of Business Venturing, 25, 5, 2010, 481-492.

Istituto per la competitività, Rapporto I-Com 2013 sull'innovazione energetica, 2013.

Levinthal, D., Adaptation on rugged landscapes. Management Science, 7, 1997.

Netval, X Rapporto Netval sulla valorizzazione della ricerca Pubblica Italiana, Seminiamo Ricerca per Raccogliere Innovazione, 2013.

Nicholls-Nixon, C., Jasinski, D., The blurring of industry boundaries-an explanatory model applied to telecommunications. Industrial and Corporate Change, 4, 1995, 755–768.

OECD, ICT Applications for the Smart Grid: Opportunities and Policy Implications, OECD Digital Economy Papers, N.190, OECD Publishing. http//dx.doi.org/10.1787/5k9h2q8v9bnl-en., 2012.

Schaltegger, S., Wagner, M., Sustainable entrepreneurship and sustainability innovation. Business Strategy and Environment, 20, 2011, 222-237.

Investigating the Effect of Social Media on Trust Building in Customer-Supplier Relationships

Fabio Calefato, Filippo Lanubile, Nicole Novielli
Università degli Studi di Bari Aldo Moro,
Dipartimento di Informatica, Via Orabona 4, 70125 Bari (BA)
{fabio.calefato, filippo.lanubile, nicole.novielli}@uniba.it

Abstract. *Trust is a concept that has been widely studied in e-commerce since it represents a key issue in building successful customer-supplier relationships. In this sense, social software represents a powerful channel for establishing a direct communication with customers. As a consequence, companies are now investing in social media for building their social digital brand and strengthening relationships with their customers. In this paper, we present the preliminary results of an ongoing research aimed at investigating the role of social media in the process of trust building, with particular attention to the case of small-medium enterprises. Our findings show that social media contribute to build more affective trust than traditional websites, suggesting that social media have the potential to enhance the business of SMEs other than large companies, by fostering the affective commitment of customers.*

Keywords: Trust Building, Social Media, Empirical Studies, Human Factors, Social Computing.

1. Introduzione

Negli ultimi anni, l'uso del Web ha influenzato significativamente la comunicazione interpersonale, grazie alla diffusione capillare dei social media. Twitter, Facebook e LinkedIn sono solo alcune tra la più famose piattaforme di social networking online e sempre più numerose sono le aziende che investono oggi in strategie di marketing, sfruttando i social media per diffondere l'immagine associata al proprio marchio nonché per costruire e rafforzare le relazioni di fiducia con i clienti nuovi e abituali [Blanchard, 2011]. Ciò vale anche per le piccole-medie imprese (PMI), che possono trarre beneficio dalla viralità della comunicazione su social media come versione contemporanea del tradizionale passaparola [Tvesovat e Kouznetsov, 2011].

Costruire rapporti di fiducia con i propri clienti e fornire un'immagine affidabile della propria azienda sono entrambe questioni cruciali in ambito commerciale [Büttner e Göritz, 2008], soprattutto nel campo dell'e-commerce, in

cui la dimensione della comunicazione faccia a faccia è inesistente. Il motivo del successo dei social media come luogo per l'implementazione di strategie di marketing ha origine dal successo dei classici paradigmi di marketing. Le piattaforme di social networking, infatti, restituiscono alle aziende operanti sul Web la possibilità di gestire i rapporti con i clienti in modo personale e personalizzato. Questo è fondamentale poiché la ricerca ha dimostrato come spesso la fiducia sia stabilita nei confronti di uno specifico venditore, piuttosto che rispetto al marchio [Doney e Cannon, 1995]. In questo senso, il personale gioca un ruolo chiave nel relazionarsi con i clienti quando è in grado di fare ricorso a strategie di persuasione basate su argomentazione sia razionale sia emotiva [Petty e Cacioppo, 1986]. I social media consentono ai fornitori di realizzare questo comportamento in un ambiente virtuale in modo da fornire ai clienti la percezione di un'azienda più raggiungibile e preoccupata delle loro necessità [Blanchard, 2011]. Una PMI che miri ad attrarre nuovi clienti fornendo un'immagine affidabile di sé e a fidelizzare i propri clienti di lunga data, dovrebbe prendere in considerazione la possibilità di sfruttare il canale affettivo come strategia chiave di persuasione. In tal senso, i social media offrono la possibilità di implementare forme di interazione che simulano i tradizionali scambi faccia a faccia con i clienti, migliorando l'immagine dell'azienda in termini di affidabilità e benevolenza [Blanchard, 2011].

In questo articolo descriviamo i risultati preliminari di uno studio in corso volto a comprendere il ruolo dei social media nella costruzione di un rapporto basato sulla fiducia tra PMI e consumatori. In particolare, focalizziamo l'attenzione sull'impatto delle diverse modalità web (sito web tradizionale vs *fanpage* su social media) sulla fiducia affettiva, ossia la fiducia costruita sulla base della valutazione di argomenti e informazioni che fanno leva sulla sfera emotiva.

2. Stato dell'arte e motivazione

La fiducia è stata oggetto di studio in numerosi ambiti applicativi [Rusman et al, 2010], dalle scienze cognitive [Castelfranchi e Falcone, 2000] all'economia [Doney e Cannon, 1995], e più recentemente nell'ingegneria del software [Al-Ani e Redmiles, 2009; Schumann et al, 2012]. Secondo Hung et al. [2004], il riporre fiducia in qualcuno implica la convinzione che il fiduciario si comporterà secondo le nostre attese. Per quanto riguarda il marketing, diverse sono le definizioni di fiducia elaborate dai ricercatori. In particolare, il processo di costruzione della fiducia si sviluppa principalmente lungo diverse dimensioni che possono essere identificati come antecedenti della fiducia o '*trust antecedent*' [Rusman et al, 2010], vale a dire, le proprietà del fiduciario che innescano la valutazione cognitiva finalizzata a misurarne l'affidabilità. Coerentemente con il punto di vista dei ricercatori nel campo dell'e-commerce [Büttner e Göritz, 2008] [Dooney e Cannon, 1995], il modello adottato in questo studio è un adattamento del *Tripod Model* [Mayer et al, 1995] che definisce l'attendibilità di una persona o di un'organizzazione in termini di *abilità*, *benevolenza* e *integrità.*

L'*abilità* consiste nella capacità del fiduciario di adempiere ad un compito o un'obbligazione, rispondere a una richiesta, fornire una risposta competente. È valutata in funzione delle capacità professionali dell'altro, delle sue conoscenze e delle sue competenze specifiche. Ne sono un esempio, la descrizione dell'esperienza lavorativa di un professionista nel suo curriculum o le informazioni sulla storia professionale dell'azienda e del suo staff, solitamente fornite a scopo informativo sul sito web della società. La *benevolenza* si riferisce a caratteristiche al livello di cortesia, atteggiamento positivo, disponibilità, intenzione di condividere le informazioni e/o le risorse, disponibilità ad aiutare, gentilezza e ricettività. Una persona o azienda che soddisfi questi requisiti è solitamente percepita come incline ad accontentare le esigenze della propria clientela e a fornire supporto in caso di necessità. Per *integrità* si intende un insieme di norme morali e le caratteristiche del fiduciario considerate auspicabili dalla morale comune come, per esempio, integrità, onestà, correttezza, lealtà e discrezione.

Per quanto riguarda il dominio commerciale, McKnight et al. [1998] estendono questo modello includendo la quarta dimensione di *prevedibilità* del comportamento del fiduciario. La prevedibilità è un concetto legato a quello di responsabilità (*accountability*) enunciato da Rusman et al [2010], ossia la capacità di una persona (l'azienda, in questo caso) di soddisfare le aspettative dell'acquirente, nel dominio di riferimento, in termini di affidabilità e consistenza rispetto agli standard dichiarati. Questa quarta dimensione è rilevante in ambito e-commerce [Büttner e Göritz, 2008] e dunque considerata nel presente studio.

Rispetto al processo di costruzione della fiducia, è possibile evidenziare la differenza tra fiducia cognitiva (*cognitive trust*) e fiducia affettiva (*affective trust*) [McAllister, 1995]. La fiducia cognitiva si basa su meccanismi di valutazione razionali delle caratteristiche del fiduciario oggettivamente riconducibili al dominio di riferimento, come ad esempio le competenze professionali. Tale valutazione viene solitamente combinata con il processo di ponderazione dei vantaggi e dei rischi derivanti dal fidarsi dell'altro. Al contrario, la fiducia affettiva si manifesta in presenza di legami affettivi e di sincera preoccupazione per il benessere altrui [Hung et al, 2004].

Schumann et al. [2012] definiscono un'operazionalizzazione di questa distinzione mappando gli antecedenti della fiducia rispetto alle due dimensioni, cognitiva e affettiva. Manterremo tale corrispondenza in questo articolo, in riferimento a tutti e quattro gli antecedenti discussi finora, come rappresentato in Fig. 1. In particolare, l'abilità e la prevedibilità sono valutate per mezzo di elaborazione cognitiva delle informazioni personali e professionali del fiduciario, rilevanti rispetto al dominio di riferimento. Al tempo stesso, la valutazione lungo la dimensione affettiva porta alla costruzione della fiducia in presenza di elementi che supportano la percezione dell'altrui benevolenza e integrità.

In questo studio analizziamo come la valutazione di informazioni a disposizione sul fiduciario sia influenzata dai diversi mezzi di comunicazione online, con riferimento alle dimensioni cognitiva e affettiva. In particolare, ci riferiamo ai tradizionali siti web come *information oriented*, ossia come orientati principalmente a veicolare contenuti informativi. Al contrario, i social media permettono di osservare il comportamento di un'azienda mentre interagisce con

i propri clienti. Accedendo a un profilo aziendale su social network, i clienti possono valutare come lo staff risponde a critiche o commenti negativi, verificare che si prenda attivamente cura della relazione con i clienti, e valutare quanto il personale e i titolari si dimostrino aperti e disponibili a essere contattati. Come evidenziato da ricerche sul marketing [Blanchard, 2011], i social media possono essere utilizzati per simulare con successo una relazione uno-a-uno e amplificare la percezione di ricevere cure e attenzioni personali dedicate, come tipicamente avviene nei mercati tradizionali in cui la componente di interazione faccia a faccia è ancora presente.

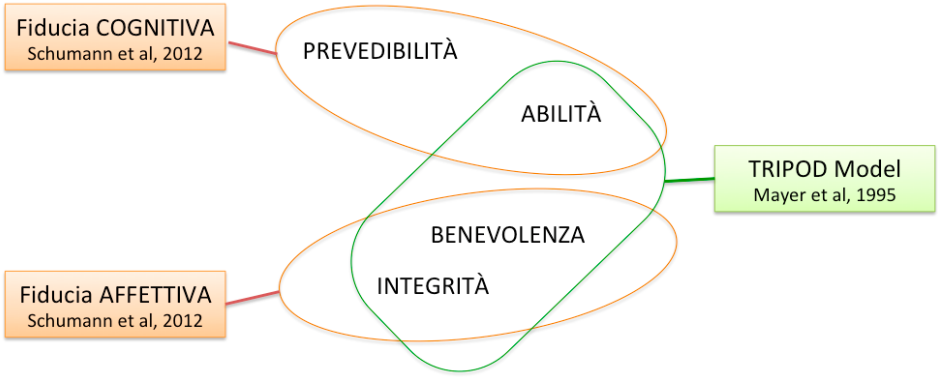


Fig.1 –Dimensione cognitiva e affettiva nel modello di costruzione della fiducia.

3.Obiettivi

3.1 Dominio

Come evidenziato da recenti ricerche, la volontà di fidarsi è inversamente proporzionale al rischio intrinseco associato al dominio di riferimento. Un esempio tipico è l'acquisto di prodotti tecnologici [de Ruyter et al, 2001] o farmaceutici [Büttner et al, 2006]. La decisione di fidarsi di un'azienda dipende anche dal rischio percepito rispetto alla classe di prodotto specifico, oltre che dalle caratteristiche del venditore e dallo specifico contesto di acquisto [Büttner e Göritz, 2008]. Nel nostro esperimento, abbiamo selezionato due piccoli ristoranti poiché per l'utente comune è possibile valutare l'affidabilità di simili aziende in conformità a semplici criteri basati buon senso senza che la valutazione richieda alcuna particolare abilità o conoscenza specifica. Le aziende A e B sono, rispettivamente, una pasticceria e un ristorante biologico. Entrambe hanno iniziato la loro attività circa un anno fa e attuano strategie di marketing basate su una forte presenza su web e social media.

3.2 Ipotesi

Nel nostro modello, assumiamo che la fiducia cognitiva sia determinata dalla valutazione positiva delle informazioni professionali, solitamente acquisite attraverso il sito web dell'azienda. Al contrario, i social media possono contribuire allo sviluppo di un'interazione con il cliente più informale e amichevole, rispondendo ai commenti e alle domande degli utenti sulle community online (ad esempio, sulla fanpage dell'azienda su Facebook). Tale potenziale dei social media può essere impiegato a favore dall'aumento della fiducia lungo la dimensione affettiva. Ipotizziamo, dunque, che la possibilità di monitorare e rispondere adeguatamente e in tempo reale al feedback della *community* online offra all'azienda la possibilità di migliorare la propria immagine in termini di benevolenza e integrità. In altre parole, ipotizziamo che i siti web tradizionali, orientati a comunicare contenuti informativi, e i social media, strutturati in modo da favorire l'interazione diretta con i clienti, possano avere un impatto diverso, rispettivamente, sulla fiducia cognitiva e affettiva.

Di conseguenza, formuliamo due ipotesi che, rispetto alla creazione di rapporti di fiducia tra azienda e nuovi potenziali clienti sulla base della prima impressione ricavata dall'osservazione del profilo aziendale online:

H_aff *– I social media favoriscono la* ***fiducia affettiva*** *più dei siti web tradizionali***.**

H_cog *– I siti web tradizionali favoriscono la* ***fiducia cognitiva*** *più dei social media*.

4. Lo studio

Lo studio è progettato secondo uno schema 2 x 2 (Tab. 1). La modalità Web (*Sito web tradizionale* vs *Profilo su social media*) costituisce l'unica variabile indipendente, mentre l'azienda (*A* vs *B*) è considerata semplicemente un fattore di blocco poiché quello che intendiamo indagare è l'effetto della modalità web sulla fiducia. L'esperimento ha coinvolto 19 studenti di laurea specialistica in informatica (16 maschi, età media = 25 anni). Su una scala Likert a 4 punti, tutti i partecipanti tranne uno hanno dichiarato di utilizzare Facebook su base giornaliera (media = 3.74). La maggior parte di loro ha inoltre dichiarato di utilizzare spesso Internet per acquisti (media = 2.58). E' dunque ragionevole supporre che il campione presenti un livello omogeno di familiarità con il Web e con il meccanismo di valutazione dell' affidabilità di un'azienda attraverso la sua presenza online tipico dell'e-commerce. L'ordine di visualizzazione di sito web e pagina Facebook per ognuno delle due aziende è stato randomizzato per evitare condizionamenti imputabili alla sequenza di presentazione.

Ogni partecipante ha valutato l'affidabilità percepita di entrambe le aziende rispetto a una delle due possibili combinazioni (Tab. 1). La fiducia è valutata da ciascuno dei partecipanti rispetto a tre degli antecedenti nel modello: abilità, prevedibilità e benevolenza. Coerentemente con la letteratura [Schumann et al, 2012], la valutazione dell'integrità è esclusa poiché è un antecedente peculiare di relazioni a lungo termine. Nel nostro studio, invece, i

partecipanti non conoscevano le aziende prima dell'esperimento: sarebbe impossibile fornire una valutazione dell'integrità in un simile contesto in cui la non conoscenza delle aziende costituisce un prerequisito per la partecipazione all'esperimento.

Per la valutazione, è stato definito un questionario [Calefato et al, 2013] ottenuto integrando le linee guida e gli elementi inclusi nel questionario di studi sulla fiducia basata sulla prima impressione [Büttner e Göritz, 2008; Rusman et al 2010, de Ruyter et al, 2001]. Per ogni antecedente è stato definito un insieme specifico di domande: 7 per abilità, 3 per prevedibilità, 11 per benevolenza (questionario basato su scala Likert da 1 a 5). La valutazione è associata alle due dimensioni cognitiva e affettiva, secondo quanto discusso nella Sezione 2.

Modalità web x Azienda	*A*	*B*
Sito web tradizionale	Gruppo 1	Gruppo 2
Profilo su social media	Gruppo 2	Gruppo 1

Tab.1 – Progettazione dello studio

4.4 Procedimento

L'esperimento si è svolto in ambiente controllato: tutti i partecipanti sono stati coinvolti simultaneamente sotto la supervisione dello stesso sperimentatore. Lo sperimentatore ha introdotto e spiegato l'esperimento ai partecipanti contemporaneamente, illustrando lo scenario e fornendo istruzioni dettagliate sulla sequenza di esecuzione delle attività. Lo sperimentatore è rimasto in laboratorio durante lo svolgimento dell'esperimento per garantire che i partecipanti non interagissero né scambiassero opinioni sulle due aziende, nonché per rispondere alle loro domande durante l'esperimento. Ogni studente ha lavorato in modo indipendente, accedendo alla procedura sperimentale attraverso un'apposita interfaccia guidata sul web.

All'inizio dell'esperimento, i partecipanti hanno risposto a un breve questionario introduttivo volto a valutare il loro grado di familiarità con le tecnologie web, i social network e l'e-commerce. In seguito, è stato loro introdotto il task dell'esperimento ossia la scelta di una delle due aziende proposte come fornitore per l'allestimento di un buffet. Lo scenario della ristorazione è stato scelto poiché presenta un fattore di rischio più elevato rispetto alla semplice scelta di un ristorante per un pranzo individuale, poiché nello scenario riguardante il catering è in gioco una somma di denaro più consistente e che l'eventuale insuccesso di un buffet presenta necessariamente anche delle implicazioni sociali rispetto agli invitati all'evento.

Una volta presentato lo scenario, i partecipanti sono stati invitati a visualizzare e valutare i due profili aziendali attraverso il sito web dell'azienda o il suo profilo di Facebook, secondo l'ordine di presentazione casuale descritto in precedenza. I partecipanti hanno esplorato il profilo aziendale per non più di cinque minuti. Di conseguenza, ciascuna delle due fasi di valutazione è durata al massimo quindici minuti, includendo visualizzazione del profilo online e compilazione del questionario. Pre-condizione necessaria per la partecipazione

all'esperimento era la non conoscenza da parte dello studente né del sito web né del profilo Facebook di nessuna delle due aziende. Al termine dell'esperimento, i partecipanti sono stati sollecitati a esprimere pareri e opinioni in modo non strutturato nel corso di una discussione collettiva con lo sperimentatore.

5.Risultati

Le due ipotesi *H_aff* and *H_cog* sono state valutate eseguendo un'analisi multivariata della varianza (MANOVA) sulle due varabili dipendenti, ossia la fiducia affettiva e la fiducia cognitiva. Il test ha rivelato che l'interazione tra la modalità web e l'azienda ha un effetto significativo sul livello di fiducia affettiva stabilita sulla base della prima impressione ricevuta dai partecipanti osservando i profili online delle aziende (F = 4.528, p = .041), come mostrato dal grafico di interazione in Fig. 2a. Al contrario, nessun effetto significativo è stato osservato sullla fiducia cognitiva assumendo un livello di significatività pari a .05, né per l'interazione tra modalità web e azienda (F = .035, p = .853) né per gli effetti principali di modalità web (F = 1.590, p = .216) e azienda (F = .977, p = .330) (Fig. 2b).

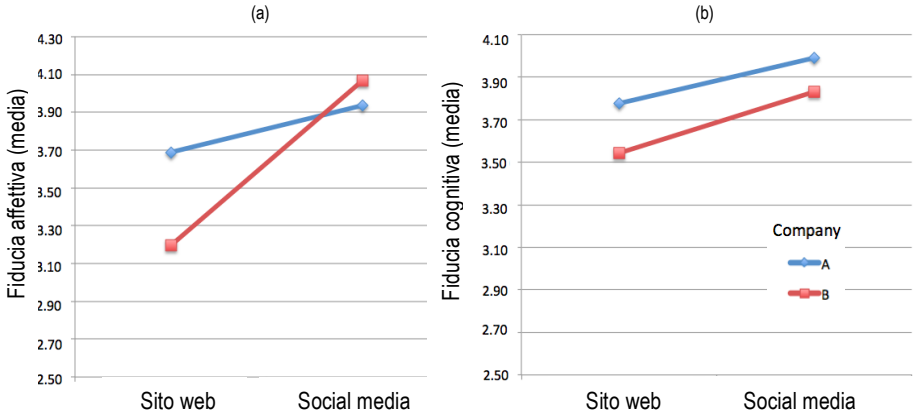


Fig. 2. Effetto dei fattori su trust affettivo (a) e cognitivo (b)

Per meglio comprendere la causa dell'interazione tra modalità web e azienda sulla fiducia affettiva, abbiamo eseguito due ANOVA separate per ognuna dei due ristoranti coinvolti nell'esperimento. I risultati sono riportati in Tab, 2 e mostrano come l'uso di social media abbia un impatto significativo sulla fiducia affettiva solo per l'azienda B (F = 22.152, p = 0.000), per la quale osserviamo un valore medio di fiducia affettiva pari a 3.20 (Dev. Std. = .42) nella modalità sito web, per poi salire a 4.07 (Dev. Std. = .38) nella modalità social media. Al contrario, per l'azienda A non riscontriamo nessuna differenza statisticamente significativa tra fiducia affettiva osservata a fronte della visualizzazione di sito web (media = 3.69, Deve. Std. = .60) e profilo Facebook (media = 3.94, Deve. Std. = .35).

Azienda	Effetto della modalità web	
	F	P
A	1.211	0.287
B	22.152	***0.000***

Tab. 2 ANOVA sulla fiducia affettiva per le due aziende

5.1 Discussione

Abbiamo testato due ipotesi, rispettivamente H_aff (*i social media esercitano un maggiore impatto sulla fiducia affettiva*) e H_cog (*i siti web tradizionali esercitano un maggiore impatto sulla fiducia cognitiva*). Rispetto alla fiducia affettiva, i risultati supportano parzialmente l'ipotesi H_aff suggerendo che effettivamente il ricorso al web marketing attraverso i social media contribuisce a creare un'immagine più aperta e benevolente dell'azienda presso i potenziali nuovi consumatori che ne visitano il profilo online, già sulla base del primo impatto. L'analisi della varianza ha evidenziato che tale impatto è statisticamente significativo solo per una delle due aziende considerate nello studio. L'interazione tra la modalità web (variabile indipendente) l'azienda (fattore di blocco) suggerisce la necessità di controllare, nelle future repliche dell'esperimento, l'effettiva equivalenza sia delle informazioni fornite sia della strategia di comunicazione attuata dalle due aziende.

Rispetto alla fiducia cognitiva, i risultati non supportano l'ipotesi H_cog poiché non rileviamo nessun impatto significativo da parte dei siti web delle due aziende sulla costruzione della fiducia rispetto alla dimensione cognitiva. Tuttavia, interessanti spunti di riflessione sono stati forniti dai partecipanti nel corso del colloquio collettivo e informale con lo sperimentatore, a conclusione dell'esperimento controllato. La maggior parte dei partecipanti ha infatti dichiarato che il ricorso a fanpage e community su social media, per quanto concepite per favorire l'interazione informale con gli utenti, può favorire una migliore valutazione dell'affidabilità dell'azienda anche rispetto alla dimensione cognitiva. Infatti, i partecipanti dichiarano di valutare come più competenti le aziende che deliberatamente si espongono direttamente al feedback dei clienti sui social media, poiché tale comportamento è percepito come indice di maggiore credibilità nonché come asserzione di maggiore qualità nei beni e servizi forniti.

Inoltre, la maggior parte degli studenti coinvolti ha dichiarato di individuare nella disponibilità di foto (del ristorante, dei prodotti e dello staff) un elemento fondamentale per la valutazione della competenza e professionalità dell'azienda. Ciò è coerente con la letteratura sul rapporto tra fotografie e fiducia [Riegelsberger et al, 2013] [Steinbrueck et al, 2002] e sull'effetto che l'accesso agli elementi di informazione normalmente associati alla dimensione affettiva (ad es. foto del personale) può avere anche lungo la dimensione cognitiva [Schumann et al, 2012].

6. Conclusioni e sviluppi futuri

Nel presente studio abbiamo analizzato il ruolo dei social media nella costruzione della fiducia lungo la dimensione affettiva, in un contesto in cui potenziali clienti visualizzano per la prima volta le informazioni dell'azienda visitandone i profili disponibili sul web. Abbiamo inoltre indagato il potenziale del social software per le piccole aziende che vogliano costruirsi un'immagine online basata su affidabilità e competenza.

Abbiamo osservato come, consistentemente con le ipotesi iniziali, i social media detengano il potenziale necessario per incrementare la percezione di disponibilità e apertura rispetto alle specifiche esigenze dei clienti. Includendo il ricorso ai social media nella propria strategia di web marketing, un'azienda può simulare interazioni personali e individuali attraverso gli strumenti offerti dalle più comuni piattaforme di social networking online.

I risultati di questo studio sono da considerarsi preliminari e richiedono successiva validazione attraverso future repliche dell'esperimento che coinvolgano un campione di partecipanti più ampio e vario e introducano un maggiore controllo degli elementi informativi e dell'effettiva equivalenza delle strategie di comunicazione attuate dalle aziende selezionate.

Ringraziamenti

La presente ricerca è finanziata dal progetto Intersocial, nell'ambito del Programma di Cooperazione Transfrontaliera Grecia-Italia 2007-2013. Ringraziamo tutti gli studenti che hanno partecipato allo studio.

Bibliografia

Al-Ani, B. and Redmiles, D. In Strangers We Trust? Findings of an Empirical Study of Distributed Teams. Proc. 4th Int'l Conf. on Global Software Engineering (ICGSE '09), 2009

Blanchard, O. Social Media ROI, 2011. Pearson Education.

Büttner, O. B., and Göritz, A. S. Perceived trustworthiness of online shops. Journal of Consumer Behaviour 7, 2008, 35-50

Büttner, O. B., Schulz, S. and Silberer, G. Perceived Risk and Deliberation in Retailer Choice: Consumer Behavior towards Online Pharmacies. Advances in Consumer Research, 33, 2006, 197-202.

Calefato, F., Lanubile, F., Novielli, N. A Preliminary Investigation of the Effect of Social Media on Affective Trust in Customer-Supplier Relationships. Proc. of Affective Computing and Intelligent Interaction (ACII 2013), to appear.

Castelfranchi, C. and Falcone, R. Trust is more than subjective probability: mental components and sources of trust. Proc. of the 33rd Hawaii International Conference on System Sciences, 2000.

Doney, P. M., and Cannon J. P. An Examination of the Nature of Trust in Buyer-Seller Relationships. Journal of Marketing 61, 1995, 35-51.

Hung, Y.C., Dennis, A.R., and Robert, L. Trust in Virtual Teams: Towards an Integrative Model of Trust Formation. Proc. of the 37th Hawaii International Conference on System Sciences. HICSS '37. IEEE Computer Society, Washington DC, 2004

Mayer, R. C., Davis, J. H., and Schoorman, F. D. And integrative model of organizational trust. Academy of Mngmt Review 20, 1995, 709-734.

McAllister, D.J. Affect and cognition-based trust as foundations for interpersonal cooperation in organizations. Academy of Management Journal 38, 1, 1995, 24-59.

McKnight D.H., Cummings, L.L., Chervany, N.L. Initial trust formation in new organizational relationships. Academy of Mngmt Review, 23 (3), 1998, 473-490.

Petty, R. E., and Cacioppo, J. T. Communication and Persuasion: Central and Peripheral Routes to Attitude Change, Springer, N, 1986.

Riegelsberger, J., Sasse, M.A., McCarthy, J.D. Shiny Happy People Building Trust? Photos on e-Commerce Websites and Consumer Trust. Proc. of CHI 2003, 2003.

Rusman, E., van Bruggen, J., Sloep P., and Koper R. Fostering trust in virtual project teams: Towards a design framework grounded in a TrustWorthiness Antecedents (TWAN) schema. International Journal of Human-Computer Studies 68, 2010, 834-850.

de Ruyter, K., Moorman, L. and Lemmink, J. Antecedents of Commitment and Trust in Customer-Supplier Relationship in High Technology Markets. Industrial Marketing Management, 30, Elsevier Science, 2001, 271-286.

Schumann, J., Shih, P., Redmiles, D., and Horton, G. Supporting Initial Trust in Distributed Idea Generation and Evaluation. Proc. International ACM SIGGROUP Conference on Supporting Group Work (GROUP 2012) ACM, 2012, 199-208.

Steinbrueck, U., Schaumburg, H., Duda, S., and Krueger, T., A, Picture Says More Than A Thousand Words - Photographs As Trust Builders In E-Commerce Websites. In. Proceedings of CHI2002: Extended Abstracts, 2002, 748-749.

Tvesovat, M. and Kouznetsov, A. Social Network Analysis for Startups. O'Reilly, 2011.