# IJKBO Editorial Board

## International Editorial Review Board:

## International Editorial Review Board Continued:

## IGIP Editorial:

IGIP
IGI Publishing
WWW.IGI-GLOBAL.COM

# Call for Articles

## International Journal of Knowledge-Based Organizations

***An official publication of the Information Resources Management Association***

The Editor-in-Chief of the *International Journal of Knowledge-Based Organizations* (IJKBO) would like to invite you to consider submitting a manuscript for inclusion in this scholarly journal.

An official publication of the Information Resources Management Association
INTERNATIONAL JOURNAL OF
Knowledge-Based Organizations
IGI PUBLISHING
Publisher of IT books, journals and cases since 1988
www.igi-global.com

*ISSN 2155-6393*
*eISSN 2155-6407*
*Published quarterly*

**MISSION:**
The mission of *International Journal of Knowledge-Based Organizations* (IJKBO) is to provide an international forum for organizational and governmental practitioners, researchers, information technology professionals, software developers, and vendors to exchange useful and innovative ideas within the field. This journal emphasizes the presentation and distribution of groundbreaking, original theories and concepts shaping future directions of research enabling business managers, policy makers, government officials, and decision makers to comprehend advanced techniques and new applications of information technology. IJKBO encourages exploration, exploitation, and evaluation of different principles, processes, technologies, techniques, methods, and models in sustainable knowledge ecosystems.

**COVERAGE:**
Topics to be discussed in this journal include (but are not limited to) the following:

- Behavioral sciences
- Business forecasting
- Business intelligence
- Business mathematics
- Computer science
- Decision making in social setting
- Decision making under uncertainty
- Decision making with multimedia
- Decision sciences
- Decision theory
- Economics
- Electronic commerce
- Ethical decision making
- Fussy information processing
- Geographic clusters
- Grid analysis
- Group decision making and software
- Informatics
- Information technology
- Intelligent agents technologies
- Intelligent organizations
- Knowledge discovery in databases
- Knowledge engineering
- Knowledge-based analysis
- Knowledge-based assets
- Knowledge-based benchmarking
- Knowledge-based competition
- Knowledge-based decisions
- Knowledge-based development
- Knowledge-based economy
- Knowledge-based enterprises
- Knowledge-based performance management
- Knowledge-based processes
- Knowledge-based production management
- Knowledge-based resources
- Knowledge-based services
- Knowledge-based society
- Knowledge-based strategy
- Knowledge-based systems
- Knowledge-based ventures
- Knowledge-based workflow
- Management information systems
- Mathematical optimization
- Mathematics of decision sciences
- Methods of decision making
- Morphological analysis
- Multiple criteria decision making
- Network analysis
- Ontological Engineering
- Operations research
- Perspectives of decision making
- Political decision making
- Psychology
- Public decision making
- Risk management
- Robust decision making
- Scenario analysis
- Sensitivity analysis
- Sequential decision making
- Simulation methods
- Sociology
- Statistical decision
- Styles of decision making
- System analysis
- System theory
- Technological forecasting
- Virtual organizations
- Web-based decision making

All submissions should be emailed to:
John Wang, Editor-in-Chief
E-mail: journalswang@gmail.com

**Ideas for special theme issues may be submitted to the Editor-in-Chief.**

**Please recommend this publication to your librarian. For a convenient easy-to-use library recommendation form, please visit: http://www.igi-global.com/ijkbo**

# International Journal of Knowledge-Based Organizations

April-June 2013, Vol. 3, No. 2

## Table of Contents

### Research Articles

*International Journal of Knowledge-Based Organizations* is listed or indexed in: Bacon's Media Directory; Cabell's Directories; DBLP; Google Scholar; INSPEC; JournalTOCs; MediaFinder; ProQuest Advanced Technologies & Aerospace Journals; ProQuest Computer Science Journals; ProQuest Illustrata: Technology; ProQuest SciTech Journals; ProQuest Technology Journals; The Standard Periodical Directory; Ulrich's Periodicals Directory

# Attacks on Confidentiality of Communications Between Stranger Organizations

*Mikaël Ates, Entr'ouvert, Paris, France*

*Gianluca Lax, Department of Computer Science, Electronics, Mathematics and Transportation (DIMET), University Mediterranea of Reggio Calabria, Reggio Calabria, Italy*

## ABSTRACT

*Knowledge has become the main intangible asset of many organizations. Two organizations that have had no previous contact or relationship are defined strangers. When two stranger organizations enter into a relationship, knowledge plays a very critical role since each party has to disclose its own knowledge to achieve knowledge from the other party. In this paper, the authors study the confidentiality of communications between stranger organizations, showing that even when strong authentication algorithms, like RSA, are exploited, no guarantee that the communication is confidential can be given. This study is surely useful to keep in mind the limitations concerning the confidentiality whenever stranger organizations are involved.*

*Keywords:* *Communication Between Strangers, Confidentiality, Privacy, Stranger Organizations, Trust*

## INTRODUCTION

In the industrial era, organizations were based on the physical work done by individuals in factories and, for this reason, they are referred as job-based organizations. Nowadays, the idea of workers holding a job is replaced by that of workers bringing knowledge to the organization, knowledge understood as capability of solving a problem, identifying a threat, analyzing a scenario, and so on. In this case, we refer to *knowledge-based organizations* (KBOs) (Lindgren Stenmark, & Ljungberg, 2003). Knowledge has become the main intangible asset of KBOs and plays a very critical role when two organizations enter into a relationship, since each interlocutor has to disclose its own knowledge in order to achieve knowledge from the other party (Allison & Strangwick, 2008). The above issue in also more critical whenever the two organizations are *strangers*, that is, they have no a priori information about their interlocutor and they have to take decisions (Heikkinen, Matuszewski, & Hammainen, 2008). This typically happens when organizations are in open environments, such as the Internet or ubiquitous and pervasive environments, where the perceived risk is high

DOI: 10.4018/ijkbo.2013040101

(Cunningham, Gerlach, Harper, & Kellogg, 2008). In this case, the environment is marked by the following three characteristics:

- **Absence of Identifiers (*C1*):** Claiming an identity to the interlocutor is useless;
- **Insecure Channel (*C2*):** The communication takes place over an insecure channel;
- **Absence of a Single Trusted Party (*C3*):** There is no single authority able to ensure security services (e.g., confidentiality).

*C1* derives directly from the definition of strangers: indeed, since organizations are strangers, any identifier name is unknown to the interlocutor. *C2* is due to the fact that we are in open environments so that an attacker can sniff, modify, intercept, kill, re-route, delay, and reorder messages (Srinivasulu, Nagaraju, Kumar, & Rao, 2009). *C3* models the fact that it is possible to have an external party trusted by two or more organizations but it is unrealistic to assume the presence of a third party trusted by all organizations.

It is expected that in the future there will be a lot of stranger organizations that will have the necessity of entering into a relationship for business reasons. In this paper, we analyze the relationship between stranger organizations and, in particular, the concepts of *trust* and *confidentiality*. Indeed, relationships between strangers in an open environment rely on trust, considered as a state in which an entity accepts to enter into a relationship with another entity, expecting to reach a goal (McKnight & Chervany, 2001). Confidentiality, usually defined as a service used to keep secret the content of a communication from all but those authorized to have access (Menezes, Vanstone, & Oorschot, 1996), may be requested for relationships. In this study, we formalize the task in which two entities enter into a relationship. Then, we analyze the confidentiality of their communication and we show possible threats.

We observe that the topic of communication between strangers is very relevant also in many other application contexts, like P2P systems, C2C e-commerce, ad-hoc networks, privacy preserving, authentication, trust negotiation, unlinkability, only to cite some examples (some of them will be discussed in the related work section). It is worth noting that we do not introduce here a new technology or a solution to the problem of confidentiality since, as we explain in the paper, no solution exists. Conversely, the contributions of our study can be summarized as follows:

- We define relationships between strangers in an open environment. Even though many works of the literature deal with this topic, they always assume less restrictive hypotheses (see the related work section). To the best of our knowledge, our paper is the first one that formalizes and analyzes such relationships over an environment constrained by the characteristics previously introduced.
- We highlight possible threats and attacks on confidentiality in this scenario. In particular, we explain why, also the use of strong cryptographic algorithms, like RSA, cannot guarantee confidentiality. New technologies often bring some illegitimate expectations and this study is useful to keep in mind their limitations concerning the confidentiality paradigm. The open challenge arisen from our study is finding which are the less restrictive hypotheses to be assumed in order that the problem can have a solution.
- We prove by model checking that the attacks presented in this paper succeed. The model here built could be used to validate future protocols assuming less restrictive hypotheses or to highlight their possible flaws.

The rest of the paper is organized as follows. We start with the description of the reference scenario by a running example. Then, we give a formal definition of the scenario and we introduce the definitions of trusted and authorized (for confidentiality) entities. We follow this with an analysis about possible attacks on the confidentiality of the communication between two stranger organizations. the efficacy of such attacks is proved by model checking. Then, we compare our study with the state of the art and, finally, we draw our conclusions and discuss future work.

## RUNNING EXAMPLE

In order to help the reader in understanding the scenario, we here describe a simple situation where the two stranger entities are organizations which need to enter into a relationship. In the relationship between strangers, the trust of each entity in another one is restricted to the information provided by the counterpart and gathered through the communication interface. For the sake of presentation, we call *shower* the entity providing the information and *consumer* the entity gathering the information. Such information can be *certified*, when is guaranteed (e.g., digitally signed) by a third party trusted by the consumer, or *uncertified*, when is self-asserted by the shower or guaranteed by a third party that is not trusted by the consumer. Strangers can establish a relationship whenever the gathered information is enough to satisfy all trust requirements. Such requirements can be viewed as the access control rules of an access control policy (Damiani, De Capitani di Vimercati, & Samarati, 2005) having the relationship as goal.

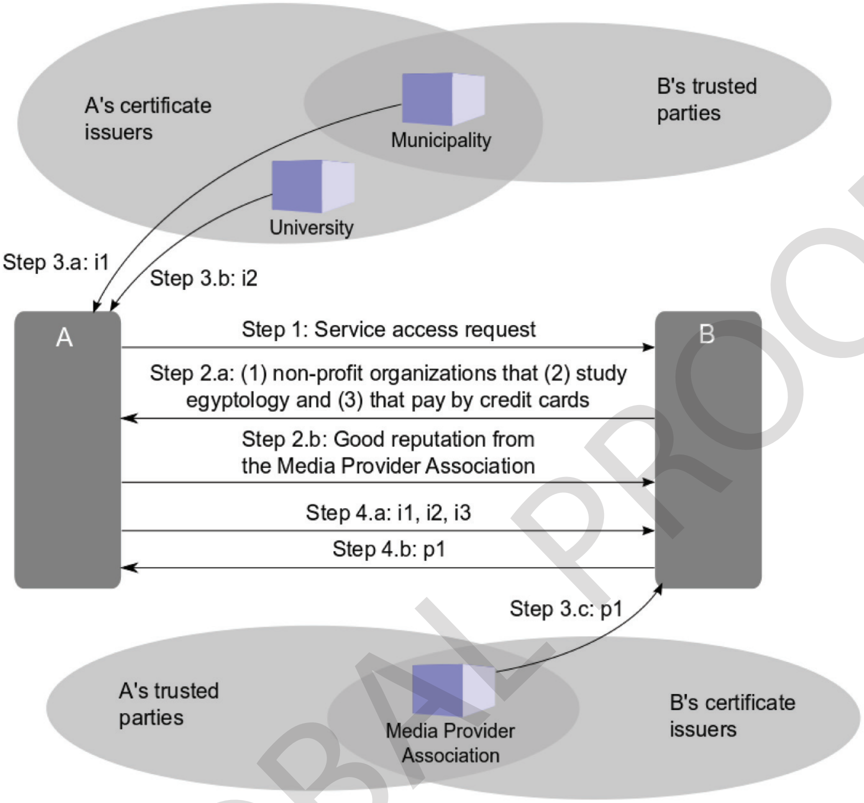
In Figure 1, we report an example of the information exchange between the two entities aimed to establish a relationship. We call *transaction* such information exchange. In our example, the first entity, say *A*, is a school and the second entity, say *B*, is a service provider specialized in documentary. The two entities *A* and *B* are strangers (i.e., this is the first time they enter into contact and they have no previous knowledge of each other) and they make use of the Internet to communicate among them.

The school *A* is interested in downloading a documentary on Pharaohs so that it contacts the service provider *B* (Step 1). The provider replies that it can provide this service only to (1) non-profit organizations that (2) study Egyptology and (3) that pay by credit cards. These requirements are the rules of the access control policy of the provider. The school replies that it is able to satisfy these requests. However, since a credit card number is necessary for the payment, the school requires that the provider can prove to have a good reputation from the *Media Provider Association* (*MPA*). Note that since this aspect is not crucial for understanding the paper, we do not discuss about the quantitative meaning of the term *good*. This step (Step 2) can be seen as a negotiation because each party accepts to disclose some information but requires gathering other information. Clearly, negotiation may end in a failure if no agreement is found and in this case the transaction aborts. It is worth noting that, during a transaction, each entity behaves both as a shower and as a consumer. Indeed, the school is a shower when proves to be a non-profit organization and a consumer when receives the reputation of the service provider.

In Step 3, each entity collects the information to be presented. The school has a digital certificate, say $i_1$, signed by the municipality proving that it is a non-profit organization and that is stored in its own secured data warehouse (Rifaie, Kianmehr, Alhajj, & Ridley, 2009). Moreover, the school has to contact the university to obtain a digitally signed declaration, say $i_2$, that it is a partner in a project on Egyptology (Steps 3.a and 3.b). Finally, the school prepares a personal declaration, say $i_3$, stating that the payment will be done by credit card. Meanwhile, the provider has obtained the reputation certificate $p_1$ from the Media Provider Association. In Step 4, the two entities exchange the information above. Observe that, the information contained in $i_1$ is certified since this certificate is issued by the municipality that is trusted by both the entities. Also $p_1$ is certified since MPA is trusted by the consumer. On the contrary, $i_2$ and $i_3$ contain uncertified information. Indeed, the former is in a certified issued by a party (the university) which is not trusted by *B*, the latter is self-asserted by *A*.

In pervasive environments where actors are anonymous, it is necessary that at least a part of the gathered information be certified. For this reason, overlapping between certifi-

*Figure 1. The transaction between A and B*


A's certificate issuers
B's trusted parties
Municipality
University
Step 3.a: i1
Step 3.b: i2
A
B
Step 1: Service access request
Step 2.a: (1) non-profit organizations that (2) study egyptology and (3) that pay by credit cards
Step 2.b: Good reputation from the Media Provider Association
Step 4.a: i1, i2, i3
Step 4.b: p1
Step 3.c: p1
A's trusted parties
Media Provider Association
B's certificate issuers

cate issuers and trusted third parties (like the municipality in the example) is also necessary to establish relationships between strangers. Concerning this aspect, we realistically assume that certificates have the following properties. The first property is the *certificate possession proof*, which means that a user can prove to possess a certificate. For instance, think of a X.509 certificate that contains a public key used by the counterpart to cipher a challenge (Housley, Ford, Polk, & Solo, 1999; Housley & Homan, 1999) that can be deciphered only knowing the secret key. Another example of possession proof is an anonymous credential (Camenisch & Lysyanskaya, 2001), where a signature value is proven using a proof-of-knowledge protocol (Fiat & Shamir, 1986). The second characteristic is that certificates can be issued *runtime* by the certifier, as done for $i_2$ in the example above. Clearly, it is also possible that some certificates are issued and used in multiple negotiations without having been reissued (think about $i_1$ in the example).

## MODELLING THE RELATIONSHIP BETWEEN STRANGERS

In this section, we formally define the scenario described above and we focus on the confidentiality of the relationship. Consider given a set $E$ of stranger entities and let $a$ and $b$ be two entities belonging to $E$ that want to enter into a relationship. Let $I_b$ be the information that $b$ sends to $a$ in the transaction. We have that $a$ accepts to enter into a relationship with $b$ when the condition

$$I_b \supseteq T_a$$

is verified, where $T_a$, called *trust threshold*, is the set of information that *a* requires to enter into a relationship with its interlocutor (i.e., the access control policy of *a*).

Observe that $I_b$ can contain some identifiers. For example, *a* could require that its interlocutor is VAT registered (in order to invoice a service) so that the interlocutor sends its VAT number. A VAT Number is clearly an identifier, but it is used only to verify that the interlocutor satisfies an access control rule. Characteristic *C1* given in the introduction implies that the identifying information does not increase the level of trust in the counterpart, since entities are strangers. Then, we can state that only a part of the information related to the interlocutor (e.g., having a VAT number) is meaningful and this part does not include identifying information (e.g., the VAT number).

We denote by *mean(·)* the primitive which extracts the meaningful information from the information gathered from an entity. For example, in the case above, *mean("My VAT number is xxx")* returns only *"I have a VAT number"*. Thus, we can rewrite the above condition as

$$mean(I_b) \supseteq T_a.$$

Since meaningful information can thus be common to multiple entities (e.g., all entities having a VAT number), we deduce that a trust relationship can be established with any entity *x* member of the set of entities able to satisfy its access control policy. We can define such a set.

- **Definition:** Given a set E of stranger entities, we say that x is a trusted entity for a, with a,x $\in$ E, iff x is able to provide a set of information $I_x$ such that mean($I_x$)$\supseteq$ $T_a$.

In our running example, the set of entities trusted by the provider is composed of all the non-profit organizations that study Egyptology and pay by credit cards.

We can repeat the same reasoning also for the interlocutor *b*, obtaining that a trust relationship between the strangers *a* and *b* is achieved if both the interlocutors reach the trust threshold of the counterpart, that is:

$$mean(I_b) \supseteq T_a$$

and

$$mean(I_a) \supseteq T_b.$$

A part of the information exchanged between *a* and *b* can be certified and included in one or multiple certificates provided by a single or multiple trusted third parties. We denote by $C_b \supseteq I_b$ the set of the certified information that *b* sends to *a*. Observe that certification ensures information integrity. Conversely, the integrity of the uncertified information exchanged over a hostile channel cannot be verified.

The entities *authorized for confidentiality* (*authorized entities*, for short) are expected to keep secret the data for which confidentiality is desired. For example, in a confidential communication between two parties, the two interlocutors are the only entities authorized for confidentiality and the use of a communication protocol guaranteeing confidentiality (e.g., the SSL protocol - Freier, Karlton, & Kocher, 1996) ensures this requirement.

When two interlocutors cannot authenticate each other, the most general protocol to guarantee confidentiality consists in an unauthenticated secret key exchange, where the key is used later to cipher the communication. For example, if RSA is used, one of the two entities gives its certificate containing a public key and the other entity uses this public key to cipher a random key that will be used by both the entities for symmetric key encryption.

Since the two entities *a* and *b* are strangers, it is necessary that all the certificates used by *a* to satisfy the access control policy of *b* could be proved to be referred to *a*, and vice versa. It is worth noting that the expression

”all certificates are referred to *a*” means that, from the point of view of *b*, all certificates can be referred to the same interlocutor. For this purpose, each certificate contains a value called *key confirmation value*. When the interlocutor receives certificates with the same key confirmation value, it can be sure that they are referred to the same entity. In order to be more concrete, the key confirmation value could be the public key of *a,* so that all the certificates used by *a* have to contain the same public key (i.e., that of *a*). In this way, *a* is the only entity able to decipher the secret key encrypted with its public key. Then, the inclusion of the public key in every certificate makes *b* certain that the entity providing the public key is the same entity which presents the certificates.

Now, among all certificates in $C_b$, we distinguish that containing the key confirmation (for *b*) $k_b$ from that not containing such a confirmation. We denote these two sets by $C_b^{K_b^+}$ and $C_b^-$, respectively. We are ready to give the definition of authorized entity.

- **Definition:** Given a set E of stranger entities and a transaction between two entities a and b, we say that x is an authorized (for confidentiality) entity by a, with a,b,x $\in$ E, iff x is able to provide the same set of certified information extracted by mean( $C_b^{K_b^+}$ ).

In our running example, the set of entities authorized by the provider *B* is the set of the entities which can prove to be non-profit organizations. The presence of at least a (public key) certificate is necessary in order to exchange the secret key that will be used to cipher the transmission.

It is clear that the best case (from the security point of view) is when all information is certified and contains the key confirmation. In this case, we have that the set of trusted entities coincides with that of authorized entities and that a confidential relationship between two strangers *a* and *b* requires that:

$$mean(C_b^{K_b^+}) \supseteq T_a$$

And

$$mean(C_a^{K_a^+}) \supseteq T_b.$$

The notation used in the paper is summarized in Table 1. In the next section, we will show that, contrary to what might be expected, also in this best case some security threats occur.

## ATTACKS ON CONFIDENTIALITY

With reference to the scenario described so far, consider again the entities *a* and *b* and suppose that each entity is authorized by the other one. Let $i \in E$ be a dishonest entity that wants to compromise the confidentiality of the communication between *a* and *b*. In this section, we will show that if *i* is authorized by both *a* and *b*, then *i* is a threat. In particular, we describe a successful attack done by *i* on the key exchange protocol based on RSA (Rivest, Shamir, & Adleman, 1978).

Let us start with the protocol description. We use the notation introduced in Ryan and Schneider (2001) to describe the protocols. The standard key exchange protocol starts with *a* that sends a first certificate containing its public key, say $PK_a$, to *b*. Now *b* replies with a random value $N_b$ encrypted by $PK_a$ that will be used by *a* as a symmetric cipher session key for the communications with *b*:

$a \rightarrow b$: $PK_a$ a sends its public key to b

$b \rightarrow a$: $\{N_b\}_{PKa}$ b sends a nonce to a

The confidentiality is ensured if *b* is sure that $PK_a$ is the key sent from *a* and if *a* is sure that *b* has generated $N_b$ and encrypted it with $PK_a$. We decompose the protocol in what *a* and *b* send and receive:

a: send($PK_a$)
b: receive($PK_a$’)

*Table 1. Notation*

| Symbol | Description |
|---|---|
| $I_x$ | Information provided by the entity *x* |
| $T_x$ | Trust level threshold at which the entity *x* accepts to enter into a relationship |
| *mean(·)* | Primitive which extracts the meaningful information to trust a stranger |
| $C_x$ | Certified information provided by the entity *x* |
| $k_x$ | Key confirmation value for the entity *x* |
| $C_x^{k^+}$ | Certified information provided by *x* containing the key confirmation value *k* |
| $C_x^{-}$ | Certified information provided by *x* not containing any key confirmation value |

```
b: send({N_b}_PKa')_
a: receive({N_b'}_PKa'')
```

*a* can only decrypt the received message if $PK_a$'= $PK_a$. The confidentiality is ensured if *b* is sure that $PK_a$''=$PK_a$ and *a* is sure that $N_b$'= $N_b$. As a matter of fact, if *b* is sure that $PK_a$ is the key used by *a*, then it is sure that only *a* will receive $N_b$. If *a* is sure that *b* has received $PK_a$ and *b* has sent the secret encrypted with $PK_a$, then only *a* can decipher and see the secret.

Consider now the key confirmation value $k_a$ that will be included in the next certificates. According to the discussion done about the key confirmation value, $k_a$=$PK_a$. After the first exchange, in the second phase, *a* sends all the certificates (containing $k_a$) necessary to satisfy the access control policy of *b*. Such data are ciphered by $N_b$, as follows:

$a \rightarrow b$: $\left\{C_a^{PK_a^+}\right\}_{Nb}$

Now we introduce the dishonest entity *i* authorized by both *a* and *b* and we describe the attack which allows *i* to learn $N_b$ without *a* and *b* being able to detect that the secret is compromised. The protocol is flawed by a man-in-the-middle-based attack in the following way:

1. $a \rightarrow i$: $PK_a$ interception
2. $i \rightarrow b$: $PK_i$ substitution
3. b → $i$: $\{N_b\}_{PKi}$ $N_b$ is compromised
4. $i \rightarrow a$: $\{N_b\}_{PKa}$ $N_b$ is unchanged
5. $a \rightarrow i$: $\left\{C_a^{PK_a^+}\right\}_{Nb}$ interception
6. $i \rightarrow b$: $\left\{C_i^{PK_i^+}\right\}_{Nb}$ substitution
7. $b \rightarrow i$: $\left\{C_b^{PK_i^+}\right\}_{Nb}$ interception
8. $i \rightarrow b$: $\left\{C_i^{PK_a^+}\right\}_{Nb}$ substitution

In particular, in Step 2, *i* substitutes the public key of *a* with its one so that it can know the secret $N_b$. Then, in Steps 6 and 8, it sends its own certificates (containing its key confirmation value $PK_i$) producing the same meaningful certified information as that of the certificates of *a* and *b*, respectively. It is worth noting that the last operation is possible since *i* is authorized by both *a* and *b*, so that it is able to provide the necessary certified information. The result of the attack is that the confidentiality of the communication between *a* and *b* is extended also to *i*.

Even though in our study we have shown only the attack on the protocol based on RSA, its extension to the protocol based on Diffie-Hellman or other public key protocols is straightforward. We have seen that the key exchange protocol is thus sensitive to attacks whenever the interlocutors are strangers. The most relevant result is that the confidentiality of the communication between strangers can only be restricted to the set of the entities authorized for confidentiality. In practice, in our running example, if there is another entity *i* which is authorized by both *a* and *b* (i.e., it is a non-profit organizations that studies Egyptology, has a credit card number and a good reputation from the Media Provider Association), then *i* can obtain the documentary for free, either by paying the provider and getting the same amount from the school or by directly giving the school's credit card number to the provider in the case such a piece of information is not certified.

The analysis done above considers the best case from the security point of view, which occurs when all information is certified and all certificates contain the key confirmation value. However, we have to consider the possibility that a portion of the information could be uncertified, for example because it regards the entity that, thus, should self-assert it or because this information cannot be certified by nature (e.g., the beauty of a person).

In the general case, using certified information without key confirmation or uncertified information produces a differentiation of the two sets of trusted and authorized entities because there will be authorized entities that are not trusted. In this case, the confidentiality is restricted to the authorized entities and not to the trusted ones. In practice, with reference to our running example, the ciphered communications between the school and the provider can be intercepted by any non-profit provider having a good reputation from the Media Provider Association.

## PROOF OF THE ATTACK

This section is devoted to prove by model checking that the attack previously described succeeds. We use the model checker FDR (Formal Systems, 2010) where all specifications (e.g., the protocol, the properties to check, the intruder knowledge, and so on) are described by means of the process algebra Communicating Sequential Processes (CSP) (Hoare, 1978; Roscoe, Hoare, & Bird, 1997; Schneider, 1999). CSP is a mathematical framework for the description and analysis of systems which consists of components (processes) interacting via message exchange (Ryan & Schneider, 2001). Since writing specifications in CSP is not straightforward, we exploited CASPER (Lowe, 1998), an interface to describe specifications in the CAPSL language (Millen, 1996) that will be automatically rewritten in CSP (see Ryan and Schneider (2001) for an introduction of these tools). der (2001) for an introduction of these tools). For the sake of presentation, the technical detail about the specification description is provided in Appendix.

The results of this analysis are presented according to the syntax provided by Ryan and Schneider (2001). The first result of the model checking in reported in Figure 2. It is proved that if an intruder, say *Ivo,* is not authorized by the entity *a* or *b*, then the confidentiality of the communication between *a* and *b* is safe.

The second result is shown in Figure 3. Here, the intruder is provided with all the certificates necessary to become an entity authorized for confidentiality by both *a* and *b*. We obtain that in this case the confidentiality between *a* and *b* is compromised and the successful attack is displayed on Figure 4. In particular, the intruder *Ivo* has learned the shared key and the key confirmation values have not allowed *a* and *b* to detect the attack.

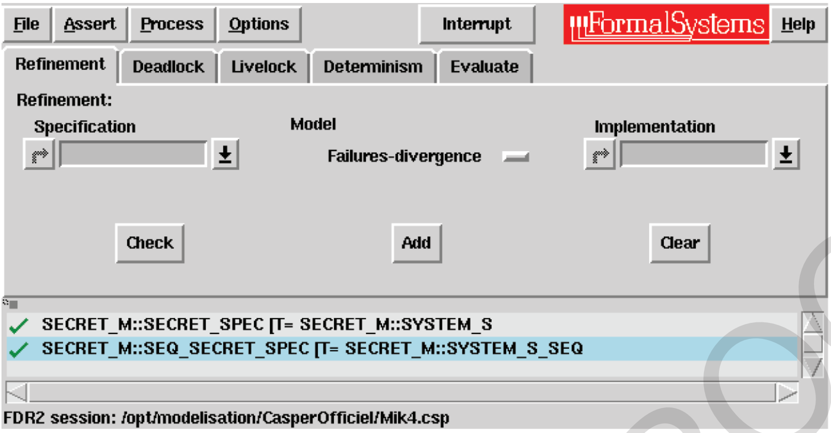
## RELATED WORK

The topic of confidentiality has received a lot of attention and has been widely studied in several application contexts.

However, it is worth noting that, as far as we know, the topic of confidentiality between stranger organizations in the scenario considered in our paper (i.e., marked by the three characteristics given in the introduction) has never been formally dealt with.

In order to highlight the importance of the topic treated in this paper, now we describe the literature regarding the communication

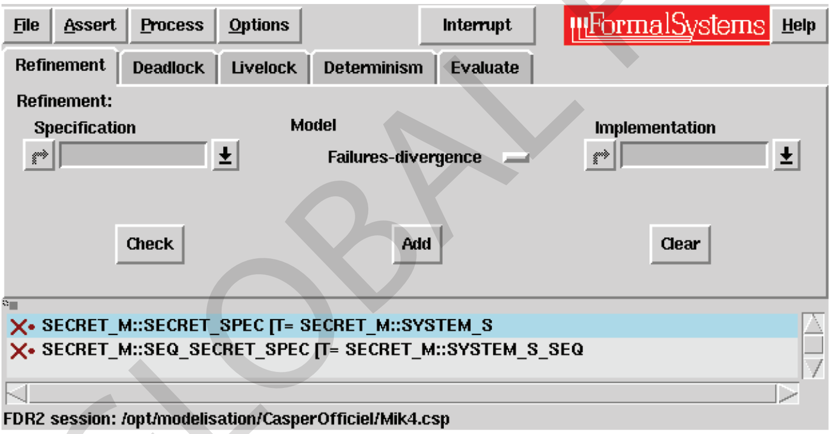
*Figure 2. Model checking of the protocol*



```
✓ SECRET_M::SECRET_SPEC [T= SECRET_M::SYSTEM_S
✓ SECRET_M::SEQ_SECRET_SPEC [T= SECRET_M::SYSTEM_S_SEQ
```

FDR2 session: /opt/modelisation/CasperOfficiel/Mik4.csp

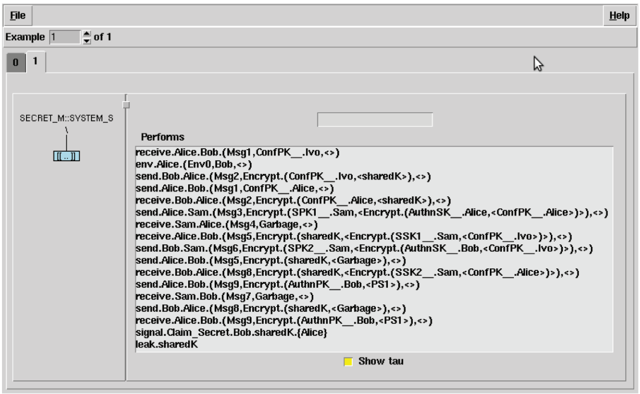
*Figure 3. Model checking with a dishonest authorized party*



```
✗ SECRET_M::SECRET_SPEC [T= SECRET_M::SYSTEM_S
✗ SECRET_M::SEQ_SECRET_SPEC [T= SECRET_M::SYSTEM_S_SEQ
```

FDR2 session: /opt/modelisation/CasperOfficiel/Mik4.csp

*Figure 4. Instantiation of the attack*



```
receive.Alice.Bob.(Msg1,ConfPK__.Ivo,<>)
env.Alice.(Env0,Bob,<>)
send.Bob.Alice.(Msg2,Encrypt.(ConfPK__.Ivo,<sharedK>),<>)
send.Alice.Bob.(Msg1,ConfPK__.Alice,<>)
receive.Bob.Alice.(Msg2,Encrypt.(ConfPK__.Alice,<sharedK>),<>)
send.Alice.Sam.(Msg3,Encrypt.(SPK1__.Sam,<Encrypt.(AuthnSK__.Alice,<ConfPK__.Alice>)>),<>)
receive.Sam.Alice.(Msg4,Garbage,<>)
receive.Alice.Bob.(Msg5,Encrypt.(sharedK,<Encrypt.(SSK1__.Sam,<ConfPK__.Ivo>)>),<>)
send.Bob.Sam.(Msg6,Encrypt.(SPK2__.Sam,<Encrypt.(AuthnSK__.Bob,<ConfPK__.Ivo>)>),<>)
send.Alice.Bob.(Msg5,Encrypt.(sharedK,<Garbage>),<>)
receive.Bob.Alice.(Msg8,Encrypt.(sharedK,<Encrypt.(SSK2__.Sam,<ConfPK__.Alice>)>),<>)
send.Alice.Bob.(Msg9,Encrypt.(AuthnPK__.Bob,<PS1>),<>)
receive.Sam.Bob.(Msg7,Garbage,<>)
send.Bob.Alice.(Msg8,Encrypt.(sharedK,<Garbage>),<>)
receive.Alice.Bob.(Msg9,Encrypt.(AuthnPK__.Bob,<PS1>),<>)
signal.Claim_Secret.Bob.sharedK.{Alice}
leak.sharedK
```

between strangers. In the relationships between strangers, trust negotiation is an intuitive and powerful means to establish gradual trust (Winsborough, Seamons, & Jones, 2000; Herzberg, Mass, Michaeli, Ravid, & Naor, 2000). This field is rich of works that aim at defining trust-negotiation properties, like the completeness and the minimal disclosure, which are studied by Yu, Ma, and Winslett (2000) and Seamons, Winslett and Yu (2001) and the access control systems, languages and algorithms, as done by Li, Winsborough, and Mitchell (2001), Winslett et al. (2002), Yu, Winslett, and Seamons (2003). Another field in which the communication between strangers is relevant is that of privacy. Nowadays, a great care must be taken to protect users' privacy to prevent technology fears (Khasawneh, Bsoul, Obeidat, & Azzam, 2010). For example, the approach of *k-anonymity* (Ciriani, De Capitani di Vimercati, Foresti, & Samarati, 2009) is exploited to preserve privacy and expects that the information gathered on a party defines a set of k entities. In Ciriani et al. (2009) where the privacy of the communication (i.e. the confidentiality) is defined, the issue of communication between strangers is very important. Indeed, having two interlocutors who exchange information under *k-anonymity* implies that they are strangers. The concept of strangers is also important in the field of unlinkability, a property deeply applied to certificates. Signature schemes based on group signature or blind signature allows making a signature not linkable between its issuing and its showing. Signature schemes allowing building certificates with an unlinkable signature have been proposed by Chaum (1983), Brands (2000), Camenisch and Lysyanskaya (2001). Unlinkability is related to the indistinguishability of an element in a set of this kind of elements (Pfitzmann & Kohntopp, 2001). It is usually expected that a transaction, where an interlocutor presents such a certificate, be not linkable to the transaction of the certificate issuing and to any other transaction where this same certificate is presented. Thus, it is expected that transactions be not linkable by certificates. The final goal is to have unlinkable transactions, which means a strong anonymity for the interlocutors which have led such transactions.

In the following sections, we present the most relevant works related to our study, which we recall focuses on relationships between strangers in an environment defined by three characteristics: absence of identifiers (*C1*), insecure channel (*C2*) and absence of a single trusted party (*C3*). Related work is subdivided according to the method exploited to face the problem of communication between strangers.

## Authentication

A first possibility is forcing an authentication mechanism in order that the interlocutors can not be considered strangers anymore. Clearly, this solution contrasts the first characteristic.

In Adams and Lloyd (1999), the term *unknown* is used interchangeably with the term *strangers*, with the meaning of *digitally unknown only*. Therein, an identity is known before the first digital contact: Entities are unknown to each other, but "Alice knows that there is a lawyer named Bob with whom she need to have a confidential conversation". Thus, in Adams and Lloyd (1999), public-key cryptography is identified as a way to solve the issue of confidentiality between strangers. The authors state that one of the driving motivations behind public-key cryptography was the inherent difficulty of enabling secure communications between strangers in a symmetric cipher environment. The binding of public-key cryptography and identity management relies on public-key certificates introduced in Loren (1978). However, with public key certificates, a stranger can only be considered as digitally unknown and a known alias is used for authentication. Our notion of *stranger* or *unknown interlocutor* implies that there is no entity authentication possible in that sense, since there is no known identity to prove. As a consequence, a public key infrastructure, defined by the NIST (National Institute of Standards and Technology, 2010) as procedures, components and facilities to bind user names to electronic keys, is inappropriate.

All the fields of the communication security domain are subjects of works on communications between strangers. For instance, the fields of ad-hoc wireless networks and peer-to-peer are rich in studies on confidentiality of communications between strangers, and usually, they rely on authentication.

For example, Aboudagga, Refaei, Eltoweissy, DaSilva, and Quisquater (2005) presented a synthesis of multiple contributions addressing this issue in ad-hoc wireless networks. Few of them give an original answer to this issue. For instance, Manweiler, Scudellari, Cancio, and Cox (2009) present an interesting concept for relationships in mobile environments not based on certified information. Strangers obtain trust from shared information issued from physical events already happened (called missed connections). However, even in this case, entities identify each other through the exchange of pre-shared information, which eludes characteristic *C1*.

## Secure Channel

Some works circumvent the rule of a very first contact between strangers over a hostile channel with the assumption of a built-in privileged channel resistant to interception. More generally, *bootstrapping a secure channel* is the term used, thus evading the second characteristic of our scenario.

Works on ad-hoc wireless networks have raised solutions relying on bootstrapping, like that of Stajano and Anderson (2000), where the idea of a side channel is introduced. In particular, they present the resurrecting duckling security policy model, which describes secure transient association of a device with multiple serialized owners. Balfanz, Smetters, Stewart and Wong (2002) proposed a pre-authentication performed over a location-limited side-channel, where the strangers are assumed to be physically close enough to create a privileged side channel, so that they can securely exchange their public keys. This subject was recently revisited by Xu and Capkun (2008) where the authors present a general framework for bootstrapping of mobile ad-hoc networks. They also give methods to select the relevant system parameters in the Random Walk models. Also Shin, Gunter, Kiyomoto, Fukushima, and Tanaka (2009) face the problem of initiating secure communication for ad-hoc network devices. They define the notion of location-limited channel and propose the use of pairing-based cryptography as a solution for bootstraps supporting ownership enforcement and key-escrow.

However, a secure channel works if both the interlocutors know that their expected interlocutor is effectively the one at the endpoint of the secure channel. It is the assumption made in the contributions above, where it is supposed that an ad-hoc network allows it. We can then finally reduce the use of bootstrapped secure channels to the need of identification of the secure channel endpoints.

## Centralization

Relying on a central authority is a way to ensure confidentiality if both the interlocutors trust this authority. Clearly, *centralization* eludes the third characteristic.

Centralized architectures were proposed for key distribution. Needham and Schroeder (1978) introduced a central server to achieve authenticated communication in computer networks. The proposed solution, which relies on conventional and public-key encryption algorithms, allows the establishment of authenticated connections and the management of authenticated mail. Also Otway and Rees (1987) describe a protocol for efficient mutual authentication which is based on a mutually trusted third party that assures both principal parties of the timeliness of the interaction without the use of clocks or double encryption. Another example of centralized solution is Kerberos (Kohl & Neuman, 1993).

With the issue of strangers, such an authority would be also in charge to determine the authorized entities for each interlocutors and then distribute the keys. However, as previously stated, such an assumption does not make sense in open environments, except in some particular case, like for instance the secret handshakes.

The secret handshake technology was simultaneously introduced in Balfanz et al. (2003) and in Biskup and Karabulut (2003). It is a powerful means to establish a secure relationship between strangers, possibly between more than two interlocutors. Members of a same group can secretly learn this membership. It means that non-members can not prove this membership and are not able to learn the membership of members. However, it is always relying on a single authority managing all the involved parties. An instance of its implementation is presented by Tsudik and Xu (2006), where a framework for secret handshake is proposed. Therein, an unauthenticated key agreement protocol is introduced and a group authority in charge of key distribution is exploited in order to address the sensitivity to a man in the middle attack.

## CONCLUSION

In the modern era, knowledge has become the main intangible asset of KBOs and plays a very critical role whenever two organizations enter into a relationship. In this paper, we have studied the trust relationship between organizations showing that, in pervasive environments, the communication between stranger organizations can not be guaranteed to be confidential. Indeed, also when strong cryptographic algorithms and the best security requirements are exploited, the communication can be intercepted by any intruder which is considered trusted by the two interlocutors. The attack allowing the intruder to break the communication confidentiality has been presented and proven by model checking. This study is thus important to keep in mind the intrinsic limitations of the considered scenario, especially when new technologies come out. The next step is finding which hypotheses have to be (hopefully) partially relaxed in order that a solution can exist. This is the open challenge arisen from our study. Currently, we are studying the possibility of slackening the third one (i.e., absence of a common trusted party). In particular, we are thinking about authorities dedicated to initiate relationship between strangers. Each authority could be specialized in a particular kind of relationship, such as commercial transactions, social relationship, and so on. They also should be in charge of key distribution.

As a future work we are focusing on the task of negotiation and on the implementation of the primitive mean($\cdot$). Indeed, in this paper, we have assumed the existence of a single access control policy disclosure. In a real scenario, usually the access control policy is dynamical and established according to other parameters, like the time or the physical location, and its disclosure can be negotiated. Concerning the primitive mean($\cdot$), we are studying a mechanism by means of which the certified information revealed to the interlocutor is exactly the required information. Therefore, we are designing an algorithm to extract the meaningful information from an existing access control policy. This is useful also to limit the amount of (unnecessary) information sent to the counterpart and also to reduce privacy concerns.

## ACKNOWLEDGMENT

The first author thanks Francesco Buccafurri and the Dipartimento di Informatica, Matematica, Elettronica e Trasporti (DIMET) and the University Mediterranea of Reggio Calabria for having received him to lead a part of his research.

## REFERENCES

Aboudagga, N., Refaei, M. T., Eltoweissy, M., DaSilva, L. A., & Quisquater, J. J. (2005). Authentication protocols for ad hoc networks: taxonomy and research issues. In *Proceedings of the 1st International Workshop on Quality of Service & Security in Wireless and Mobile Networks* (pp. 96–104). ACM.

Adams, C., & Lloyd, S. (1999). *Understanding public-key infrastructure: Concepts, standards, and deployment considerations*. Macmillan Technical Publishing.

Allison, I., & Strangwick, C. (2008). Privacy through security: Policy and practice in a small-medium enterprise. Computer Security, Privacy and Politics: Current Issues, *Challenges and Solutions*, 157-179.

Balfanz, D., Durfee, G., Shankar, N., Smetters, D., Staddon, J., & Wong, H. C. (2003). Secret handshakes from pairing-based key agreements. In *Proceedings of the 2003 IEEE Symposium on Security and Privacy (SP '03)* (p. 180). IEEE Computer Society.

Balfanz, D., Smetters, D. K., Stewart, P., & Wong, H. C. (2002). Talking to strangers: Authentication in ad-hoc wireless networks. In *Proceedings of the Symposium on Network and Distributed Systems Security (NDSS '02)*.

Biskup, J., & Karabulut, Y. (2003). Mediating between strangers: A trust management based approach. In *Proceedings of the 2nd Annual PKI Research Workshop*.

Brands, S. A. (2000). *Rethinking public key infrastructures and digital certificates: Building in privacy*. Cambridge, MA: MIT Press.

Camenisch, J., & Lysyanskaya, A. (2001). An efficient system for non-transferable anonymous credentials with optional anonymity revocation. In *Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques* (pp. 93–18). Springer- Verlag.

Chaum, D. (1983). Blind signatures for untraceable payments. In *Proceedings of Advances in Cryptology (CRYPTO '82)* (Vol. 82, pp. 199–203).

Ciriani, V., De Capitani di Vimercati, S., Foresti, S., & Samarati, P. (2009). Theory of privacy and anonymity. In *Algorithms and Theory of Computation Handbook* (2nd ed.). CRC Press.

Cunningham, L. F., Gerlach, J., Harper, M. D., & Kellogg, D. L. (2008). Perceived risk for multiple services in the consumer buying cycle. [IJISSS]. *International Journal of Information Systems in the Service Sector*, *1*(4), 33–49. doi:10.4018/jisss.2009062903.

Damiani, E., De Capitani di Vimercati, S., & Samarati, P. (2005). New paradigms for access control in open environments. In *Proceedings of the 5th IEEE International Symposium on Signal Processing and Information* (pp. 540–545).

Fiat, A., & Shamir, A. (1986). How to prove yourself: Practical solutions to identification and signature problems. In *Proceedings of Advances in Cryptology (CRYPTO '86)* (pp. 186–194). Springer- Verlag.

Formal Systems. (2010). Received May 10, 2010, from http://www.fsel.com

Freier, A., Karlton, P., & Kocher, P. (1996). The SSL protocol version 3.0. In Netscape Communications.

Heikkinen, M., Matuszewski, M., & Hammainen, H. (2008). Scenario planning for merging mobile services decision making: Mobile peer-to-peer session initiation protocol case study. [IJIDS]. *International Journal of Information and Decision Sciences*, *1*(1), 26–43. doi:10.1504/IJIDS.2008.020034.

Herzberg, A., Mass, Y., Michaeli, J., Ravid, Y., & Naor, D. (2000). Access control meets public key infrastructure, or assigning roles to strangers. In *Proceedings of the 2000 IEEE Symposium on Security and Privacy (SP '00)* (p. 2). Washington, DC: IEEE Computer Society.

Hoare, C. A. R. (1978). Communicating sequential processes. *Communications of the ACM*, *21*(8), 666–677. doi:10.1145/359576.359585.

Housley, R., Ford, W., Polk, W., & Solo, D. (1999). Internet X509 public key infrastructure certificate and CRL profile. *IETF RFC 2459*.

Housley, R., & Hoffman, P. (1999). Internet X509 public key infrastructure operational protocols: FTP and HTTP. *IETF RFC 2585*.

Khasawneh, A., Bsoul, M., Obeidat, I., & Azzam, I. A. (2010). Technology fears: A study of e-commerce loyalty perception by Jordanian customers. *International Journal of Information Systems in the Service Sector*, *2*(2), 70–77. doi:10.4018/jisss.2010040105.

Kohl, J., & Neuman, C. (1993, September). The kerberos network authentication service v5.

Li, N., Winsborough, W. H., & Mitchell, J. C. (2001). Distributed credential chain discovery in trust management: Extended abstract. In *Proceedings of the 8th ACM conference on Computer and Communications Security (CCS '01)* (pp. 156–165). ACM.

Lindgren, R., Stenmark, D., & Ljungberg, J. (2003). Rethinking competence systems for knowledge-based organizations. [EJIS]. *European Journal of Information Systems*, *12*(1), 18–29. doi:10.1057/palgrave.ejis.3000442.

Loren, M. (1978). *Toward a practical public-key cryptosystem*. Unpublished doctoral dissertation, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Lowe, G. (1998). Casper: A compiler for the analysis of security protocols. [JCS]. *Journal of Computer Security*, *6*(1-2), 53–84.

Manweiler, J., Scudellari, R., Cancio, Z., & Cox, L. P. (2009). We saw each other on the subway: secure, anonymous proximity-based missed connections. In *Proceedings of the 10th Workshop on Mobile Computing Systems and Applications (Hotmobile '09)* (pp. 1–6). New York, NY: ACM.

McKnight, D., & Chervany, N. (2001). The meanings of trust. *Trust in Cyber-Societies*. *LNAI*, *2246*, 27–54.

Menezes, A. J., Vanstone, S. A., & Oorschot, P. C. V. (1996). *Handbook of applied cryptography*. Boca Raton, FL: CRC Press, Inc. doi:10.1201/9781439821916.

Millen, J. K. (1996). Capsl: Common authentication protocol specification language. In *Proceedings of the 1996 Workshop on New Security Paradigms* (pp. 132). New York, NY: ACM.

*National Institute of Standards and Technology.* (2010). Retrieved May 10, 2010, from http: www.nist.gov

Needham, R. M., & Schroeder, M. D. (1978). Using encryption for authentication in large networks of computers. *Communications of the ACM*, *21*(12), 993–999. doi:10.1145/359657.359659.

Otway, D., & Rees, O. (1987). Efficient and timely mutual authentication. *ACM SIGOPS Operating Systems Review*, *21*(1), 8–10. doi:10.1145/24592.24594.

Pfitzmann, A., & Kohntopp, M. (2001). Anonymity, unobservability, and pseudonymity - a proposal for terminology. In. Lecture notes in computer science: Vol. 2009. *Designing privacy enhancing technologies* (pp. 1–9). Berlin / Heidelberg, Germany: Springer. doi:10.1007/3-540-44702-4_1.

Rifaie, M., Kianmehr, K., Alhajj, R., & Ridley, M. J. (2009). Data modelling for effective data warehouse architecture and design. [IJIDS]. *International Journal of Information and Decision Sciences*, *1*(3), 282–300. doi:10.1504/IJIDS.2009.027656.

Rivest, R. L., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, *21*(2), 120–126. doi:10.1145/359340.359342.

Roscoe, A. W., Hoare, C. A. R., & Bird, R. (1997). *The theory and practice of concurrency*. Upper Saddle River, NJ: Prentice Hall PTR.

Ryan, P., & Schneider, S. (2001). *The modelling and analysis of security protocols: The CSP approach*. Addison-Wesley Professional.

Schneider, S. (1999). *Concurrent and real time systems: The CSP approach*. New York, NY: John Wiley & Sons, Inc..

Seamons, K., Winslett, M., & Yu, T. (2001). Limiting the disclosure of access control policies during automated trust negotiation. In *Proceedings of the Symposium on Network and Distributed System Security (NDSS '01)*.

Shin, W., Gunter, C. A., Kiyomoto, S., Fukushima, K., & Tanaka, T. (2009). How to bootstrap security for ad-hoc network: Revisited. *IFIP Advances in Information and Communication Technology*, *1-3*, 119–131. doi:10.1007/978-3-642-01244-0_11.

Srinivasulu, P., Nagaraju, D., Kumar, P., & Rao, K. (2009). Classifying the network intrusion attacks using data mining classification methods and their performance comparison. *International Journal of Computer Science and Network Security*, *9*(6), 11–18.

Stajano, F., & Anderson, R. J. (2000). The resurrecting duckling: Security issues for ad-hoc wireless networks. In *Proceedings of the 7th International Workshop on Security Protocols* (pp. 172–194). London, UK: Springer-Verlag.

Tsudik, G., & Xu, S. (2006). A flexible framework for secret hand-shakes (multi-party anonymous and unobservable authentication). In. Lecture Notes in Computer Science: Vol. 4258. *Privacy Enhancing Technologies* (pp. 295–315). Berlin / Heidelberg, Germany: Springer. doi:10.1007/11957454_17.

Winsborough, W. H., Seamons, K. E., & Jones, V. E. (2000). Automated trust negotiation. *DARPA Information Survivability Conference and Exposition*, 1, 0088.

Winslett, M., Yu, T., Seamons, K. E., Hess, A., Jacobson, J., & Jarvis, R. … Yu, L. (2002). Negotiating trust in the web. In Internet Computing, IEEE, 6(6), 30-37.

Xu, S., & Capkun, S. (2008). Distributed and secure bootstrapping of mobile ad hoc networks: Framework and constructions. [TISSEC]. *ACM Transactions on Information and System Security*, *12*(1), 1–37. doi:10.1145/1410234.1410236.

Yu, T., Ma, X., & Winslett, M. (2000). Prunes: An efficient and complete strategy for trust negotiation over the internet. In *Proceedings of the 7th ACM Computer and Communication Security (CCS '00)* (p. 210-219). ACM Press.

Yu, T., Winslett, M., & Seamons, K. E. (2003). Supporting structured credentials and sensitive policies through interoperable strategies for automated trust negotiation. [TISSEC]. *ACM Transactions on Information and System Security*, *6*(1), 1–2. doi:10.1145/605434.605435.

*Mikaël Ates is researcher for the Entr'ouvert company. He obtained his PhD at the Université de Lyon – University of Saint-Etienne, Télécom Saint-Etienne Engineer School, DIOM Laboratory, in 2009, in Security Architectures, Applied Cryptography, Trust and Privacy. He actually pursues his research works for Entr'ouvert and visits research teams, as currently the Computer Science research team of Professor Francesco Buccafurri at the University Mediterranea of Reggio Calabria. He also gives lectures at the Télécom Saint-Etienne Engineer School on computer science. Eventually, he is active in the free software domain.*

*Gianluca Lax is an Assistant Professor of computer science in the Department of Computer Science, Electronics, Mathematics and Transportation (DIMET) at the University Mediterranea of Reggio Calabria, Italy. In 2000, he took the Laurea degree in Electronic Engineering at the University Mediterranea of Reggio Calabria. In 2005, he took the PhD degree in computer science at the University of Calabria. Since November 2005, he is Assistant Professor at the University of Reggio Calabria, Faculty of Engineering. He is also responsible of a number of computer science courses within Master courses. His research interests include data reduction, data streams, user modelling, P2P systems, e-commerce and information security. He has published in top level international journals and conference proceedings and he has served and serves as a referee for international journals and conferences. He is involved in several national and international research projects.*

# APPENDIX

## Specifications in CSP and CAPSL

In this appendix, we describe the specifications of the protocol in CSP, where each entity is defined according to the signals it receives and sends. We use the notation introduced by Ryan and Schneider (2001) to describe the protocols.

Given two entities *a* and *b*, the notation *env?b: Agent* means that *a* receives a signal triggering the communication with *b*. The definition of *a* in CSP is the following:

Initiator(a, $PK_a$) =
env?b: Agent → send.a.b.$PK_a$→
receive.b.a.$\{N_b\}_{PKa}$ →
□

$$N_b \in Nonce \qquad \begin{pmatrix} send.a.b.\left\{\left\{PK_a\right\}_{SSK(s1)}\right\}_{Nb} \rightarrow \\ receive.b.a.\left\{\left\{PK_a\right\}_{SSK(s2)}\right\}_{Nb} \rightarrow \\ Session\left(a,b,N_b\right) \end{pmatrix}$$

The definition of *b* is:

Responder(b, $N_b$) =
receive.a.b.$PK_a$→
□

$$PK_a \in PublicKey \qquad \begin{pmatrix} send.b.s.\left\{N_b\right\}_{PKa} \rightarrow \\ receive.a.b.\left\{\left\{PK_a\right\}_{SSK(s1)}\right\}_{Nb} \rightarrow \\ send.b.a.\left\{\left\{PK_a\right\}_{SSK(s2)}\right\}_{Nb} \rightarrow \\ Session\left(a,b,N_b\right) \end{pmatrix}$$

We now introduce the trusted third party. We assume that each interlocutor establishes an identity with its credential provider by a safe authentication mechanism using a dedicated key pair denoted by *AuthnPK/AuthnSK*. These keys are different from the keys employed to exchange the secret and denoted by *ConfPK/ConfSK*. Let *s* be a third party trusted by both a and b which uses two different pairs of keys denoted by *SPK1/SSK1* and *SPK2/SSK2*.

The actors and their knowledge are described as follows in Box 1.

We introduce the CASPER notation *m%v* which means that an agent receiving a message *m* stores it in the variable *v*.

Firstly, *b* receives a key and uses it to encrypt a nonce *sharedKey*:

*Box 1.*

```
INITIATOR(a,s) knows AuthnSK(a), AuthnPK(a), ConfSK(a),
ConfPK(a), SPK1(s), SPK2(s)
RESPONDER(b,s,sharedKey) knows AuthnSK(b), AuthnPK(b), SPK1(s),
SPK2(s)
SERVER(s,a,b) knows SPK1(s), SSK1(s), SPK2(s), SSK2(s),
AuthnPK(a), AuthnPK(b).
```

1. *a → b: ConfPK(a) % pkb*
2. *b → a: {sharedKey}{pkb % ConfPK(a)}.*

Then, *a* asks *s* for a credential. It gives *ConfPK(a)* to make the key included in the credential and authenticates by signing with *AuthnSK(a)*. The message is encrypted with the public key of *s*. Then *s* issues the credential, i.e., the signed key:

3. *a → s: {{ConfPK(a) % pks1}{AuthnSK(a)}} {SPK1(s)}*
4. *s → a: {pks1}{SSK1(s)} % cred1.*

The credential shown by *a* is the key confirmation for *b*. *b* must verify that the key in the credential is the same as the one previously received:

5. *a → b: {cred1% {pks1% pkb}{SSK1(s)}} {sharedKey}.*

The same protocol phase also for *b*:

6. *b → s: {pkb % pks2}{AuthnSK(b)}}{SPK2(s)}*
7. *s → b: {pks2}{SSK2(s)} % cred2*
8. *b → a: {cred2% {pks2% ConfPK(a)}{SSK2(s)}} / {sharedKey}.*

*a* expects that the shared key be only known by *b*, and vice versa for *b* about *a*. The goal is to check that *a* and *b* can detect if *sharedKey* has been compromised. The secrecy property is described as follows:

```
Secret(a, sharedKey, [b])
Secret(b, sharedKey, [a]).
```

We now introduce the intruder *Ivo* which is able to sniff, intercept, kill, re-route, delay, reorder, replay and fake messages. We declare the intruder and provide it with the knowledge of the other actors with their public keys and with the material for faking, which is a nonce and a key pair, as follows in Box 2.

As described by Figure 2, in this case the confidentiality is not compromised.

*Ivo* is now empowered with special skills in order to make it a dishonest authorized entity. It is provided with all the credentials used for key confirmations as presented in Box 3.

As described by Figure 3, in this case the confidentiality is compromised.

Starting from this full specification in CAPSL, CASPER produces a file in $CSP_M$ (i.e., the ASCII version of the CSP notation) that has been given in input to the model checker. The discussion about the results of the model checking is reported at the end of the section "PROOF OF THE ATTACK".

*Box 2.*

```
Intruder=Ivo  IntruderKnowledge= {Alice, Bob, Ivo, Sam, Nm, Au-
thnPK, AuthnSK(Ivo), ConfPK, ConfSK(Ivo), SPK1, SPK2}.
```

*Box 3.*

```
IntruderKnowledge={Alice, Bob, Ivo, Sam, Nm, AuthnPK,
AuthnSK(Ivo), ConfPK, ConfSK(Ivo), SPK1, SPK2, {ConfPK(Ivo)}
{SSK1(Sam)}, {ConfPK(Alice)}{SSK2(Sam)}}.
```

# The Role of Supportive Leadership and Job Design for Proactive Behavior and Self-Organization in Work Groups

*Annika Lantz, Department of Psychology, Uppsala University, Uppsala, Sweden, & Fritz Change AB Sweden, Stocksund, Sweden*

## ABSTRACT

*Research on group work has shown that supportive leadership helps improve the group's cooperation and social exchange in groups, which in turn influences the effects of the group work. This study develops a previous model on the relationship between job design, group processes, group initiative and self-organizational activities by including supportive leadership. The hypothesized model was tested using LISREL 8.30 (Jöreskog & Sörbom, 1993) in five different organizational contexts (two types of industry, elderly care, school and nuclear power plant) and in 104 work groups. The results are based on work task analysis (two studies) and questionnaires. The meaningfulness of the model was tested both in contexts where proactive behavior and self-organizational activities are desirable and in a context where proactive behavior can be damaging. Dimensions of job design, supportive leadership, group processes are interrelated and connected to self-organizational activities. Reflectivity and group initiative show the largest effects on self-organizational activities. Job design captured by work task analysis gives a better model fit and has a larger impact on self-organizational activities than self-assessed autonomy. Supportive leadership has an effect on group processes that in turn impact group initiative and self-organizational activities and a direct effect on group initiative as well.*

*Keywords:* *Group Processes, Innovation, Job Design, Leadership, Proactive Behavior, Self-Organization, Teams, Work Groups*

## INTRODUCTION

An innovation is the result of a knowledge exchange within the collective learning processes and creative, interdisciplinary cooperation along the entire chain from idea to product and maintenance (Kim, 2009; North & Güldenberg, 2008; Sawyer, 2007). Innovation research points to the importance of the management both creating the right conditions for employees on all levels within the organization to participate in innovation processes and of supporting and encouraging proactive behavior (Lorenz, 2004; Parker, Williams & Turner, 2006; Strauss, Griffin & Rafferty, 2009; West, Hirst, Richter & Shipton, 2004). Crant defines proactivity as "taking

DOI: 10.4018/ijkbo.2013040102

initiative in improving current circumstances or creating new ones; it involves challenging the status quo rather than passively adapting to present conditions" (Crant, 2000, p. 436). Strauss et al. (2009) ascertain that "proactive behavior is crucial in the process of innovation, influencing the transition from idea to idea implementation" (Strauss et al., 2009, p. 279).

Production work groups seldom come up with the idea for a totally new innovation themselves but they fulfill an important task in that they contribute input to how new products or services can be improved, or produced more efficiently. They also need to prepare themselves proactively so that they can eventually produce new products and services. 70% of the costs of an innovation process are incurred in production (Ehrelspiel, Kiewert & Lindemann, 1999). There is substantial evidence that work groups can be an effective organizational solution for innovative work (Brav, Andersson & Lantz, 2009; Kozlowski & Bell, 2003; West et al., 2004). Substantial research also points to the crucial importance of the context within which work groups operate and the amount of organizational support they receive in the form of group work leadership to the results they achieve (Kozlowski & Bell, 2003).

Strauss et al. (2009) distinguish between two different forms of proactive behavior: team member proactivity aiming at changing the team situation and the way the team works; and organization member proactivity that aims at changing the way the organization as a whole works. The interest in this survey lies in examining the contextual characteristics that might influence team member proactivity which is expressed in the form of self-organizational activities. Self-organizational activities are defined as a) proactively creating conditions and organizing work so that the group can handle new possibilities, problems or tasks and b) handling and mastering unexpected situations, problems or tasks (Brav et al., 2009).

In previous research (Brav et al., 2009; Lantz & Brav, 2007), we proposed a model of determinants of self-organizational activities in work groups where dimensions of job design, group processes (cooperation, social support and reflectivity) and proactive behavior, defined as group initiative, are interrelated and connected to self-organizational activities. We refer to group initiative in accordance with Frese, Garst and Fay's (2007) concept of personal initiative as a syndrome "that results in an individual taking an active and self-starting approach to work goals and tasks and in persisting in overcoming barriers and setbacks" (Fay & Frese, 2001, p. 97). The theoretical path model was tested on work groups on the shop floor in industry and received substantial but not complete support for the causal relationships (Lantz & Brav, 2007; Brav et al., 2009).

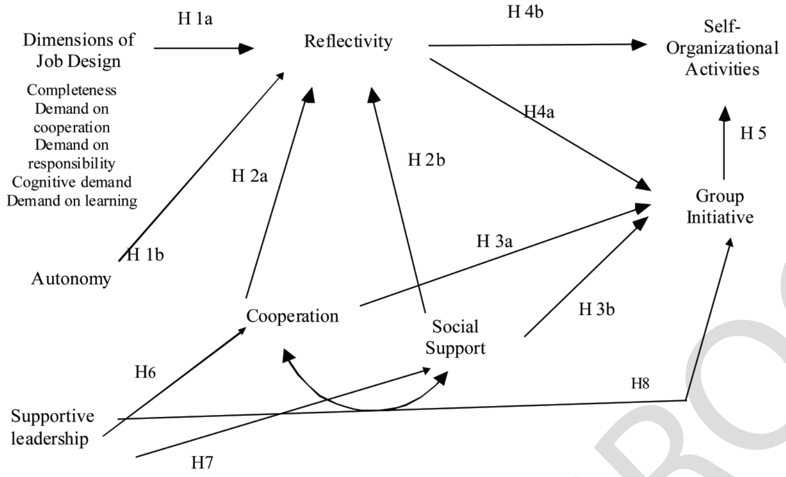
*The aim of the present study* is to extend the previous model by including supportive leadership and provide empirical support for the extended model by testing it in five different organizational contexts. The extended model is presented in Figure 1 below. The rationale behind the model is explained in the following text.

In Figure 1 above each arrow represents a hypothesis. An arrow indicates a direct effect and a double arrow represents a correlation between two variables.

## The Model

In the model, group processes mediate the relationship between dimensions of job design and group initiative as the latter presupposes reflectivity and a collective redefinition process of work. It follows the general model input - process - output model proposed by McGrath (1984), which a great deal of group research uses as a starting-point to describe and analyze group work (Kozlowski & Bell, 2003). This model describes how inputs impacts different forms of group processes, which in turn create various outputs. West et al., (2004) draw the conclusion from a literature review that there is substantial support for the assertion that "team processes provide the core driving-force for team innovation and that these processes may mediate the relationship between team inputs and innovation" (p. 91).

*Figure 1. A model for group initiative and self-organizational activities in work groups*



Taking the initiative to achieve meaningful change begins with a critical review of the prevailing conditions or situation. The group as a whole needs to reflect on goal achievement, work routines, alternative working methods, how to coordinate the work and what could be done better. The group's capacity and possibility for reflectivity, "the extent to which team members collectively reflect upon the team's objectives, strategies, and processes as well as their wider objectives" (West et al., 2004, p. 285), impact the group's readiness and motivation for change (Brav et al., 2009; West et al., 2004). An individual task is regulated in an iterative process of handling motive, goal and activity (Hacker, 2003), but collective behavior is regulated by *communication* about these issues (Tschan, 2000). The group needs to reflect and communicate on the motive, goal and activity in order to redefine the work so that it includes elements of change and development in order to achieve a new or higher goal. We use reflectivity to indirectly capture the collective redefinition process. Reflectivity creates initiative to achieve change and desire to participate in self-organizational activities. Self-organizational activities presuppose both a redefinition process of work through collective reflectivity and initiative-taking.

The scope for demonstrating group initiative and to engage in self-organizational activities is also determined by external factors such as job design. A group that cannot exert influence on how it works, or that is not permitted to participate in formulating work goals can neither take the initiative for change nor be expected to help improve work efficiency. The challenge for the management is to design the work so that it provides the conditions and the time for reflection, demands problem resolution and gives employees sufficient autonomy in order to be able to influence what is to be done and how it is done.

Activity theory (Hacker, 2003) uses the term completeness to describe if the worker can autonomously: set goals that are embedded in overall goals, prepare and plan work implementation, choose the means and interactions that benefit the work most effectively, implement, control and get feedback on his/her actions in relation to results and goal achievement. Completeness is described and evaluated through two criteria: hierarchical and sequential completeness. Sequential completeness refers to whether the individual autonomously implements a chain of activities from planning to execution and evaluation of the outcome, as these activities provide different cognitive challenges as well as feedback. Hierarchical completeness is evaluated by an examination of the means-goal relationship with regard to demand on knowledge-based and intellectual processes. Regulation demand is closely related to the concepts of autonomy and degrees of freedom. Greater autonomy and degrees of freedom increase scope for using intellectual skills, create potential for learning, redefining work and acting goal-oriented-planned.

## Hypotheses in the Original Model

A brief presentation of the theoretical support for the hypotheses (paths) in the previously tested model is given below.

Job design (assessed from the perspective of action regulation theory) is related to demand on cooperation, responsibility, cognition and learning (Lantz & Brav, 2007; Richter, Hemman & Pohlandt, 1999). A complete job will demand interaction and cooperation with others, as setting goals, the planning phase and controlling the results can rarely be done in isolation from other fellow workers and functions. Degrees of freedom and autonomy go hand in hand with taking responsibility for the work process and the results. Evaluation and feedback, as well as carrying out challenging tasks that put demand on problem solving, can be part of learning at work. Job design has been shown to be important for the individual's reflectivity and learning (Hacker, 2003) as well as for group learning and performance (Lantz & Brav, 2007). *Hypothesis 1 (H1a) in the model postulates that job design (completeness, demand on learning, responsibility, cognitive demand and cooperation) impacts collective reflectivity in work groups* and *(H1b) postulates that autonomy is related to reflectivity.*

We define cooperation as "the willful contribution of personal efforts to the completion of interdependent jobs" (Wagner, p. 152) and include coordination as such an effort. We refer to social support as socio-emotional processes such as psychological safety and trust. It is easier to behave in a cooperative manner, to share the workload and coordinate sub-tasks in an amicable climate than in a tense or non-supportive climate. Trust, psychological safety and an amicable social exchange facilitate communication about task-related issues (Edmondson, 1999). We assume that cooperation and social support are interrelated.

Brav et al., (2008) argue that a group, which has not yet established effective work routines, will be less prone to collective reflectivity than a cooperative group, since it is too busy establishing routines and getting the daily work done. Groups characterized by cooperative behavior show more reflectivity than less cooperative groups (Edmondson, 1999; Lantz & Brav, 2007). *Hypothesis 2a (H2a) postulates that cooperation will impact reflectivity.*

Social processes, such as interpersonal understanding, informal interactions, psychological safety and trust and a general amicable climate where people feel free from pressure and experience positive affect, have shown to be related to collective reflectivity and learning (Edmondson, 1999; Koslowski & Bell, 2003). A good climate allows group members to freely raise critical issues and put awkward questions without fearing potential threats or embarrassment. *Hypothesis 2b (H2b) postulates that social support will impact reflectivity.*

Constructive controversy about task-related issues is more likely to occur in a cooperative group context where members rely on everyone's willingness to contribute to reach a shared goal than in a non-cooperative group (West et al., 2004). Good cooperation can be a starting-point for finding the motivation to go beyond the stipulated task. *Hypothesis 3a (H3a) postulates that cooperation will impact group initiative.*

There is substantial evidence that a group climate characterized by openness and positive affectivity is related to idea generation, initiative-taking and innovation (Fay and Frese, 2001; West et al., 2004). It is likely that one is more motivated to expand one's work role in a setting where one feels at ease and is positive towards other group members than in a group where one is critical, dislikes or has no trust in one's fellow workers. *Hypothesis 3b (H3b) postulates that social support will impact group initiative.*

The redefinition processes presuppose collective reflectivity and are essential if the group is to form a mental task model to include proactive behavior in order to implement meaningful change. It is difficult to see how this can be done without discussions and collective reflectivity upon the state of affairs. Previous research on proactive behavior stresses the importance of dialogue and reflectivity (Hacker, 2003; West et al., 2004) for initiative-taking. *Hypothesis (H4a) postulates that reflectivity will impact group initiative.*

One needs to reflect upon the present in order to find a strategy to implement meaningful change. In line with the general argument that reflectivity is an essential part of any creative activity to change the present for the better, *Hypothesis H 4 (H4) postulates that reflectivity impacts on self-organizational activities.*

Frese, Teng, and Wijnen (1999) have shown that personal initiative is positively related to innovative work in terms of active coping strategies, having ideas and submitting suggestions for improvements, and error handling (Fay & Frese, 2001) and changes in work characteristics (Frese, Garst, & Fay, 2007). *Hypothesis 5 (H5) postulates that group initiative impacts on self-organizational activities.*

Hypothesis 1 – 5 have been previously tested (Brav et al., 2009). Hypotheses H1, H2b, H3a, H3b, H4a, and H5 were confirmed. Job design and social support have a direct impact on reflectivity and learning processes. Cooperation and social support have an impact on one output variable: group initiative. Reflectivity does not have a significant impact on group initiative, but explains the substantial amount of variance in the other output variable: self-organizational activities. Group initiative has an effect on self-organizational activities.

The results are in line with previous research. Among others, West et al. (2004), conclude that innovative groups are characterized by having a complex work task, they have many degrees in freedom in choosing the means, and they can influence the work itself and the work conditions.

## The Role of Leadership for Proactive Behavior and Self-Organizational Activities

Innovation comes from the knowledge-based exchange between employees in the organization. Research on knowledge-based organizations has shown that the management has an important task to support the interaction and knowledge exchange between employees to develop the business and innovate (Nonaka & Takeuchi, 1995; Kim, 2009). Crant (2000) and West et al. (2004) identify the management as a key aspect of the organizational context that influences proactive behavior. Empirical research on the other hand indicates that the management has a very complex task when it comes to promoting proactive behavior in work groups and there are conflicting findings regarding which type of leadership encourages initiative-taking and proactive behavior: supportive leadership and/or transformative leadership.

Research on what creates productive and efficient groups provides support for the assertion that the management plays a key supportive role in creating a positive climate and helping the groups to design good work routines and create social relationships between group members (Kozlowski & Bell, 2003; West et al., 2004). A climate of positive affectivity within an organization may provide a secure base for generating ideas and ensuring their implementation (West et al., 2004) and for ensuring quality in service (Wall, 2009).

Research on proactive behavior and leadership in organizations unexpectedly indicates that supportive leadership has no connection with proactive behavior in organizations (Frese, Teng, & Wijnen, 1999; Parker, Williams & Turner, 2006). Strauss et al. (2009) are of the opinion that supportive leadership might not be the most important leader behavior for promoting proactivity. The difference might lie in the fact that supportive leadership helps the execution of a given job, while transformational leaders motivate workers to perform beyond expectations and focus on change and improvement. The authors' survey supported their assertion that transformational leaders facilitate proactivity by increasing the employees' confidence to initiate change and by enhancing employees' commitment to the organization.

Leaders in organizations with a group-based work organization are therefore faced with a "double" task: On the one hand, they have an important task to enable both newly formed and established groups to find efficient work routines and create a good social climate by implementing supportive leadership and on the other hand they have a transformational role to challenge habitual work routines and the social exchange within the group.

This study searches for empirical support for the idea that supportive leadership has a function to create group initiative and self-organizational activities in work groups. Lantz and Laflamme (1998) showed that the more the management offered social support to the group, the better the cooperation and the social exchange within the group, which was directly related to the effects of group work. A transformational role can be important as well.

*Supportive* leadership, in the form of facilitating coordination within the group; helping the group to find efficient work routines; and providing continuous information and fast, constructive feedback on how the work is progressing, has a positive effect on the work of the group (for reviews, see for example Kozlowski & Bell; West et al., 2004). West et al., (2004) argue that the organizational environment affects innovation through information flow across the organization, participative management and decision making. Lagroue (1998) examined different types of information systems and found that the management's information support to the groups affects decision-making efficiency and decision quality. *Hypothesis (H6) postulates that supportive leadership has an impact on cooperation within the group.*

The social relationships in a work group are characterized by an exchange of both work-related and everyday items. Working in a group is not always easy. What we think and feel about each other affects our approach and our way of cooperating. Relationships to work colleagues are for most people a significant source of work satisfaction and good relationships and a good climate increase motivation and counteract the negative burden of work (Wood & West, 2010). A leadership that facilitates conflict resolution and interaction so that the group is capable of creating a good climate and can handle inevitable differences in values and opinions increases motivation and work satisfaction (Wood & West, 2010). There is substantial support for the assertion that supportive leadership has an effect on social support within the group (Wood & West, 2010; Lantz & Laflamme, 2006; 2008). *Hypothesis (H7) postulates that supportive leadership has an effect on social support within the group.*

Crant (2000) finds it probable that good leadership has a direct effect on proactive behavior without making a distinction between supportive and transformative leadership. West et al. (2004) assert that the management should both create the conditions for efficient group work and employ a leadership style that promotes proactivity. There is plenty of evidence to suggest that a management that encourages groups to question habitual routines and ingrained patterns helps to improve the groups' initiative to achieve change. Similarly, however, supportive leadership, in which the leader gives practical and informative support, encouragement and realistic feedback on the group's work can also stimulate the group's initiative. Without any direct support in previous empirical research, *the hypothesis is formulated that (H8) supportive leadership has an effect on group initiative.*

Hypotheses H6, H7 and H8 supplement the previous model and supportive leadership is assumed in the model to have an indirect impact on group initiative and self-organizational activities through group processes as well as a direct impact on group initiative.

The model and the hypotheses are illustrated in Figure 1 above.

## METHODS

To test the model's meaningfulness and generalizability, four studies were conducted in three organizational contexts where self-organizational activities are welcomed, and in a nuclear plant where readiness and alertness are crucial, but self-organizational activities can endanger safety. In the nuclear plant, the model ought not be applicable.

### Participants

The results are based on the first study conducted in four similar manufacturing enterprises (Brav et al., 2009), a manufacturing industry (Study 2), a large secondary school (Study 3), homes for the elderly with psychiatric or dementia diagnoses (Study 4) and a nuclear plant (Study 5), in Sweden. In Sweden there is a long tradition of working in groups, and in recent years a team-based organization has been introduced in many schools as well. Science can for example be taught mainly in English and different subjects can be taught parallel in a project regarding for example environmental issues. Teachers with different specialties can be responsible for the planning and execution of goal-setting, planning, teaching, and evaluation for a larger group of children who are (most of the time) divided into different sub-groups. The organizations were chosen to simply represent different sectors, organizational structures and cultures. The organizations were selected based on similar criteria:

1. Production/work planning (planning close to and in cooperation with employees)
2. Work-organizational solutions (work-groups with a shop steward/manager responsible for several workgroups)
3. Organizational support to the groups in the form of opportunities for training and competence development
4. Selection criteria for group composition (no specific or deliberate selection of group members based on demographic variables, ethnicity, personality, or differences in previous work experience)
5. Identical task instructions (to carry out a specific given, and previously stipulated task with a defined expected outcome).

The groups were selected based on the following inclusion criteria:

1. Regular meetings at least once a week with opportunities for spontaneous discussions of issues chosen by the group members
2. The group could decide which work routines would be the most effective
3. The group was allowed to organize its daily work
4. Administrative tasks and different responsibilities were rotated among group members
5. All members worked day time or the same shift and had the same principal work task
6. Group composition had been stable for at least six months
7. In Studies 1 and 2: The group practiced work task rotation

The final study groups comprised groups that met the above criteria. The number of selected groups varied between 55% (Study 4) to 88% (Study 2) of all work groups.

These groups were asked to participate and all group members received a letter describing the study as well as ethical aspects. Further information was given when the survey was administered and in all studies except for in the nuclear plant, this was done by the research team. At the nuclear plant a contact person conveyed the same information. Participation was voluntary and the surveys were completed in work time. The response rate varied between 73% (Study 5) to 96% (Study 2). Gender composition varied between male dominated groups (Studies 1 and 2) to female dominated groups (Study 4). The organizational age varied between an average of seven years (Study 2) to 14 years (Study 5). The questionnaires were coded and only the code numbers were matched with the work groups in order to ensure confidentiality. In Table 1 below the different study groups are described.

*Table 1. Participants in the different studies*

| | *Study1. Manufacturing industry 1* | *Study 2 Manufacturing industry* | **Study 3 School** | **Study 4 Care for the elderly** | **Study 5 Nuclear plant** |
|---|---|---|---|---|---|
| Work groups, N, (out of in total N), see selection criteria | 31 (42) | 36 (41) | 12 (16) | 10 (18) | 15 (21) |
| Numbers of participants N, (out of in total N), see selection criteria | 171 (239) | 189 (218) | 88 (114) | 74 (172) | 122 (189) |
| Number of participants N, (out of participants in selected groups) and response rate % | 162 (171) = 95% | 182 (189) = 96% | 82 (88) = 93% | 68 (74) = 92% | 89 (122) = 73% |
| Range of work group' size and Mean size | 3 - 11. Mean: 5 | 3 - 9. Mean: 6 | 3 - 8. Mean: 6 | 4 - 11. Mean 7 | 6 - 9. Mean 7 |
| Percentage of women and men | 17% women and 83% men | 29% women and 71% men | 56% women and 44% men | 94% women and 6% men | 7% women and 93% men |
| Average organizational age | 10 years | 7 years | 8 years | 6 years | 14 years |

Work task analysis was conducted by means of observation and observational interviews in the manufacturing industries (Studies 1 and 2). It was not possible to carry out work task analysis in all studies for a variety of reasons. At the introductory meeting, the teachers took a vote against the procedure. Their argument was that they did not feel comfortable with their work being observed and they regarded it as too complicated a procedure. The head for the elderly care unit decided it would take too much work time for the staff, and involve too much administrative time for the supervisors since all patients (or their families) had to give consent for ethical reasons. In the nuclear plant the research team was not allowed on the premises for security regulations. Therefore, in two studies the results are based on work task analysis of dimensions of job design and in all studies on a survey measurement of autonomy. This makes it possible in Studies 1 and 2 to compare the results based on work task analysis of five dimensions of job design with the results based on a survey measurement of autonomy.

## Measures

### Work Task Analysis

The REBA-instrument (in German: Rechnergestüttztes Dialogvervfahren zur psychologischen Bewertung von Arbeitsinhalten) is intended for the design and analysis of work content and job design (Pohlandt et al., 2007; Richter et al., 1999). It consists of a heterogeneous set of 22 variables that can be grouped into five theoretically and empirically related dimensions (completeness, demand on cooperation, responsibility, cognition and learning). Completeness is measured as the organizational and technological conditions determining the sequential and hierarchical completeness of work. Demand on responsibility is measured as responsibility in terms of morally and legally specified li-ability and joint responsibility for performance outcome. Cooperation is measured in terms of forms of cooperation, amount of cooperation and content of communication. Demand on cognition is measured as participation in complex planning processes and demand on problem solving. Demand on learning is measured as continuous learning demands, the use of formal training and previous work experience. (Lantz & Brav, 2007; Richter et al., 1999).

The data is obtained by observing and putting additional questions to a trained worker carrying out his/her work. The overall task is described in the number of sub-tasks and the duration of each task is noted. Twelve variables are evaluated with regard to the overall task and ten are evaluated for each sub-task respectively. The different sub-tasks were weighted according to the time it took to carry out the specific task in relation to the total work time. The scales were standardized and each overall task receives a value based on the weighted task profiles (ten variables) and the evaluation of the overall task (twelve variables). Each individual is matched with a profile for his/her job and it should be noticed that some individuals could have the same job. In Studies 1 and 2, and when the group is the unit of analysis, the results are based on the mean values of the jobs carried out in each work group. The procedure of basing the analysis on the mean value of the job may be justified, as the mean value closely resembles the value of the principal work task (Study 1 $r = .94$, $p <.01$; Study 2 $r = .89$, $p <.01$). For a full description of items and scales, see Lantz and Brav (2007).

Observer reliability was attained by using the handbook for REBA work task analyses, by using two independent observers for the analyses, and in Study 1 additional training procedures (observing and analyzing work tasks not included in the study under the supervision of an expert) and supervision by an expert. The initial inter-rater agreement was in Study 1 93% and in Study 2 89%. The discrepancies between observers (no more than one scale step in any evaluation) were further analyzed and discussed, then subjected to renewed assessment to reach absolute agreement.

## Survey Measure

Autonomy was captured by a set of six items frequently used and originally developed by Campion, Medsker & Higgs (1993) for example: "Most work-related decisions are made by the members of my work group rather than my manager".

Supportive leadership was measured by a set of six items originally developed by Lantz & Laflamme (1996) and the items have been used in earlier studies (Lantz & Laflamme, 1998 a, b). The items capture group members' perception of the support (House, 1981) given by immediate superior/supervisor, for example "We get the encouragement and support we need from our immediate superior/supervisor"; "Our immediate superior/supervisor gives us the feedback we need to know whether we are doing a good job", or "Our immediate superior/supervisor provides us with the information on conditions at the workplace we need to carry out our work task".

The six items measuring cooperation were originally designed by Campion et al. (1993), and Lantz and Laflamme (1996). They were chosen to capture workload sharing, work-task related information and communication about task-related issues in general, for example: "Everyone in our team does his/her fair share of work".

The five items measuring social support have been developed and used in earlier studies (Campion et al., 1993, Edmundson, 1999; Lantz & Laflamme, 1996) and were chosen to capture social support, psychological safety and trust: "If you make a mistake in this team, it is often held against you" (reversed).

The six items measuring Reflectivity were constructed by Edmundson (1999) and Matsson (2001) for example: "In our team, someone always makes sure that we stop to reflect on the team's work process". With the exception of three, all items capturing cooperation, social support and reflectivity were further tested by Lantz and Brav (2007).

Group initiative was measured using a set of six items developed by Frese et al., (1997), for example "I search for a solution immediately when something goes wrong". The items were transformed from individual to group level, i.e. "We search for a solution immediately when something goes wrong". To compare different survey measurements of proactive behavior: role breadth self-efficacy, (RBSE), proactive personality (PP), personal initiative (PI) and taking charge (TC), as well as to compare interviews of PI with the survey measurement, an exploratory study was conducted (Skog, 2008). Each person was interviewed to judge his/her personal initiative in accordance with the procedure suggested by Frese et al. (2007) and 55 individuals representing different professions took part in an extensive data collection in order to minimize priming. The four measures of proactive behavior all showed good internal consistency (ranging from .71–.93). The inter-correlations between survey measures of RBSE, PP, PI and TC varied between .63 to .77. Factor analysis was conducted to determine if the items represented separate constructs. A one-factor solution showed much better fit than two to four factor solutions. The correlation between interview and survey measurements of PI was .58. It was concluded that the different measurements seem to represent one construct, and as the interview and survey measurement of PI do not provide significantly different data, the use of a survey measurement can be justified.

Self-organizational activities were measured with six items first tested by North, Friedrich and Lantz (2006), an example: "In our group we have initiated change of the framework and prerequisites (conditions) for our work in order to work in the most effective way".

All items were measured on seven-point Likert-type scale with 1 for "strongly disagree" and 7 for "strongly agree". In Study 4 the response alternatives 1 and 2 as well as 6 and 7 were merged due to few responses in the extreme positions.

A Confirmatory Factor Analysis was conducted to determine whether the items of the seven scales represented separate constructs. The results show a model fit that was regarded as sufficient $\chi^2$ (454) =1136, RMSEA = .06 (see Medsker, Williams, & Holahan, 1994; Schumacker & Lomax, 2004). More importantly, the seven-factor model showed better RMSEA and fit indices than the one- to six-factor solutions.

Correlation coefficients were calculated by Pearson's formula. Internal consistency values (Cronbach's alfa) are all satisfactory and varying in the different samples between: job design .89 - .91; autonomy .59 - .82; supportive leadership .74 - .89; cooperation .79 - .92; social support .87 - .91; reflectivity .77 - .84; group initiative .85 - .94 and, self-organizational activities .84 - .88. As the dimensions of job design are highly inter-correlated in Studies 1 (.91) and 2 (.89), as well as in earlier research (Lantz & Brav, 2007; Pohlandt et al., 2007), *the dimensions are in the further analysis treated as one measurement of job design*.

To test the reliability in the group measures the intra-class coefficients ICC1 and ICC2 were computed by one-way ANOVAS (Bliese, 2000). The results varied between indexes and studies. With two exceptions (Study 5, the nuclear plant, and indexes autonomy and self-organizational activities) ICC1 showed significant F-ratios over 1 for all indexes in all studies, and significant F-ratios have previously been used in research to justify aggregation (Richter et al., 2005). All indexes (but Study 5, and indexes autonomy and self-organizational activities) reached the recommended ICC2-value of .50 or above. The conclusion was that it is justifiable to aggregate individual data to group data for Studies 1-4.

## RESULTS

Individual data could be aggregated to group data in Studies 1 - 4, but the number of groups was small in Studies 3, 4 and 5. In order to make comparison between the samples possible, the hypothesized path model (Figure 1) was tested on the group as well as the individual as the unit of analysis in Studies 1 and 2, and in Studies 3 – 5 on the individual as the unit of analysis due to the small number of work groups. It is argued, that if the results based on individual and group data in Study 1 and 2 are similar, and if the results based on individual data of the different studies are similar, it can be justified to test the model on data from the merged Studies 1 – 4, and with the group as the unit of analysis.

## Analysis of the Linear Structural Relations Between Main Variables in the Different Studies

The hypothesized path model (Figure 1) was tested using LISREL 8.30 (Jöreskog & Sörbom, 1993) and employing maximum likelihood estimation on the covariance matrix. As the reliability estimates of the manifest variables affect the parameters in the model, the error variances of the manifest variables were calculated using the reliability estimates (Cronbach Alpha), see Jöreskog & Sörbom, 1993, pp. 37-38. This procedure allows an analysis of the linear structural relations among the latent rather than the manifest variables. Model fit was evaluated by using $\chi^2$ tests first, but as sample size affects the $\chi^2$ value, the Root Mean Square Error of Approximation (RMSEA), where ≤ .05 indicates good fit (Schumacker & Lomax, 2004), and Comparative Fit Index (CFI), where ≥ .90 is considered indicative of good fit (Medsker et al., 1994), Normed Fit Index (NFI) and Non-Normed Fit Index (NNFI) where ≥ .90 is considered indicative of good fit were used as well. In small samples RMSEA ≤ .10 is acceptable, as well as lower NFI, NNFI- and CFI- values (Tabachnick & Fidell, 2007).

It was expected that the model should not be applicable to a setting where the work task is highly regulated and self-organizational activities are not welcomed, and this was confirmed in Study 5 (N=89): $\chi^2$ (8) =54.47, *p* = 00, RMSEA = .32.

The results show that the theoretical received substantial support in Studies 1- 4 and with both the individual and the group as the unit of analysis (Study 1 and 2). Model fit varies from very good to acceptable ((RMSEA: 0.000 (Study 2, the group as the unit of analysis and job design captured by work task analysis) – .07 (all samples merged, the individual as the unit of analysis and autonomy captured with a survey); CFI varied between: 1.0 - .98; NFI: 1.0 - .91, and NNFI varied between: 1.0 - .95).

The inter-correlation between cooperation and social support varied between .29 and .78 and were in all studies significant. Hypothesis 1a is confirmed in Studies 1 and 2. Hypothesis 1b is confirmed in all studies except for Study 2 (the group as the unit of analysis). Hypothesis 2a is confirmed only in two studies (studies 2 and 3). Hypothesis 2b is confirmed in Study 1 (individuals) and in Studies 2 – 4. Hypothesis 3a and 3b are confirmed in all studies. Hypothesis 4a is confirmed in Studies 1 and 2 (individuals and autonomy), and Studies 3 and 4. Hypothesis 4b is confirmed in Studies 1, 2, and 4. Hypothesis 5 is confirmed in all studies but Study 1 (Job Design and groups) where there is a tendency for group initiative to impact on self-organizational activities. Hypothesis 6, 7 and 8 are confirmed in all studies and with both the individual and the group as unit of analysis.

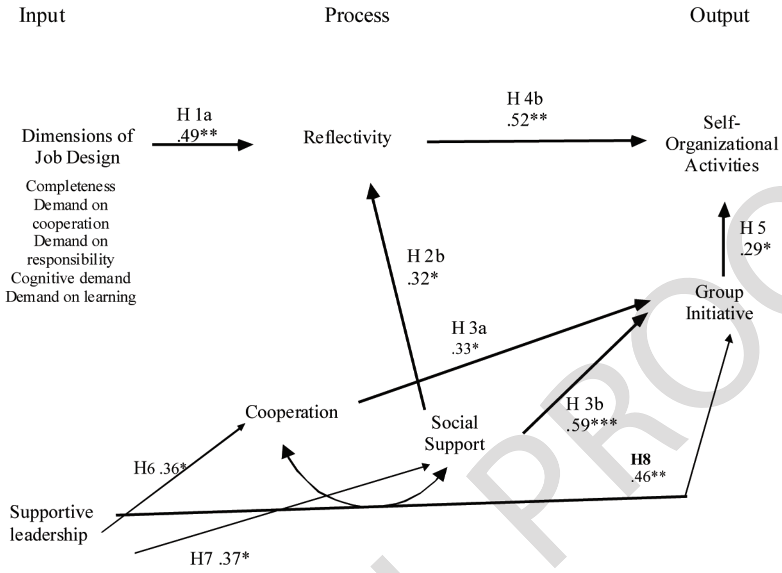
*In conclusion:* It can be justified to test the hypothesized path model (Figure 1) on merged data from Study 1- 4, with the group as the unit of analysis and with autonomy captured with a survey (N = 104). Further it can be justified to test the hypothesized path model (Figure 1) on merged data from Study 1- 2, with the group as the unit of analysis and with job design captured with work task analysis (N = 67).

## Main Results: The Relations Between Supportive Leadership, Job Design, Group Processes, Group Initiative and Self-Organizational Activities in Work Groups

### *A Model Based on Job Design Captured with Work Task Analysis and Tested on 67 Groups*

The results are shown in Figure 2 below. The final model confirmed all but two of the hypothesized paths in Figure 1. H1a (job design will be positively related to reflectivity), H2b (social support will be positively related to reflectivity), H3a (cooperation will be positively related to group initiative), H3b (social support will

*Figure 2. Final model of relations between dimensions of job design, supportive leadership, group processes, group initiative and self-organizational activities in work groups*



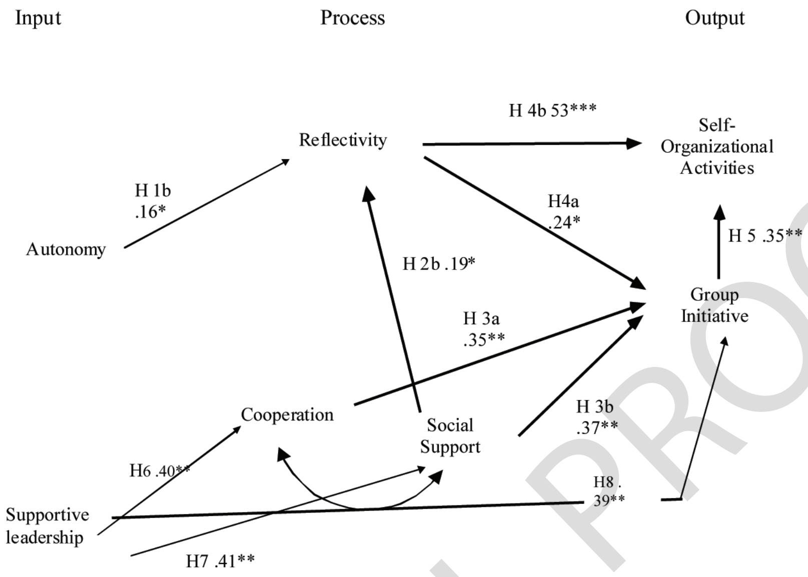
be positively related to group initiative), H4b (reflectivity will be positively related to self-organizational activities), H5 (group initiative will be positively related to self-organizational activities), H6 (supportive leadership will be positively related to cooperation), H7 (supportive leadership will be positively related to social support), and H8 (supportive leadership will be positively related to group initiative) were all supported. H2a (cooperation will be positively related to reflectivity) and H4a (reflectivity will be positively related to group initiative) had to be rejected. All path coefficients (partial regression) are statistically significant at $p < .05$. The model showed good fit to the data ($\chi^2$ (8) = 11.09, $p$ =.42, RMSEA =.02, and CFI = 0.99, NFI = .98, NNFI = .98)

### A Model Based on Autonomy Captured with a Survey and Tested on 104 Work Groups

The final model was obtained by omitting the insignificant path between cooperation and reflectivity and rerunning the model. The results are shown in Figure 3.

The final model confirmed all but one of the hypothesized paths in Figure 1. The final model provides substantial, but not complete support for the theoretical model. H1b (autonomy will be positively related to reflectivity), H2b (social support will be positively related to reflectivity), H3a (cooperation will be positively related to group initiative), H3b (social support will be positively related to group initiative), H4a (reflectivity will be positively related to group initiative), H4b (reflectivity will be positively related to self-organizational activities), H5 (group initiative will be positively related to self-organizational activities), H6 (supportive leadership will be positively related to cooperation), H7 (supportive leadership will be positively related to social support), and H8 (supportive leadership will be positively related to group initiative) were all supported. H2a (cooperation will be positively related to reflectivity) had to be rejected. Cooperation is significantly correlated with reflectivity at p < .05, but not strongly enough for a significant path coefficient. All path coefficients (partial regression) are statistically significant at p < .05. The model showed good

*Figure 3. Final model of relations between autonomy, supportive leadership, group processes, group initiative, and self-organizational activities in work groups*



fit to the data ($\chi^2$ (8) = 15.31, *p* =.46, RMSEA =.04, and CFI = 0.99, NFI = .98, NNFI = .98)

The impact of job design (work task analysis) on reflectivity is larger than that of autonomy (survey measurement). Reflectivity and group initiative show the largest effects on self-organizational activities.

## Discussion

Previous research on leadership and proactive behavior has shown that supportive leadership does not help to improve employees' proactive behavior (Strauss et al., 2009). This study has examined proactive behavior as an effect of group work. Research on group work has shown that supportive leadership does help to improve the group's cooperation and social exchange in groups, which in turn influences the effects of the group work (Lantz & Laflamme, 1998). There is substantial support for the assertion that groups can improve their proactive behavior and thereby participate in development and change work (West et al., 2004). Lantz and Brav (2007) and Brav et al. (2009) found support for the assertion that group processes influence group initiative and self-organizational activities.

The aim of the study was to develop a previous model on the relationship between job design, group processes, group initiative and self-organizational activities by including supportive leadership. The assumption was that supportive leadership benefits the development of good group processes and that it is therefore a meaningful aspect of leadership to create efficient work groups.

The hypothesized model was tested using LISREL 8.30 (Jöreskog & Sörbom, 1993). The model was tested in five different organizational contexts (two types of industry, elderly care, school and nuclear power plant) and in 104 work groups. The meaningfulness of the model was tested both in contexts where proactive behavior and self-organizational activities are desirable and in a context where proactive behavior can be damaging (nuclear plant).

A key result is that the model gains good support in contexts where proactive behavior is desirable, but not in the nuclear plant - a model which is applicable in contexts where it theoretically shouldn't be is meaningless. The results based on the individual as the unit of analysis can be regarded as valid for work groups as the reliability for a group measurement was acceptable and the SEM-analyses show similar results.

Another key result is that there is empirical support for an extended model on the relationship between job design, supportive leadership, group processes, group initiative and self-organizational activities. The model received substantial, but not complete support. There is support for all hypotheses if we consider the results of all part-studies, but the results are not entirely consistent as the effect magnitude (path coefficients) varies and not all hypotheses receive support in all studies. When data were merged from the four studies there was adequate support for all but one of the hypotheses (cooperation will be positively related to reflectivity), and there was significant correlations between the variables, i.e. there is a connection between these, but no (causal) direct effects. Job design and autonomy have a positive impact on reflectivity. Cooperation and social support are correlated to each other. Social support affects reflectivity, which in turn influences group initiative and self-organizational activities. Group initiative influences self-organizational activities. Supportive leadership affects cooperation, social support and group initiative.

The group composition regarding gender differs between industry, school and care for the elderly and the results indicate that gender does not have an impact on the model.

The results support earlier studies (Brav et al., 2009; Lantz & Brav, 2007) in that the original model is also supported in the various part-studies and when the data was merged.

In an industrial context the model shows better fit when the input-variable is captured as dimensions of job design rather than autonomy. Further, in such a setting, dimensions of job design have a larger impact on reflectivity than autonomy. The model gets substantial support also when the input-variable is assessed as autonomy, but in such a model the results are not consistent as to which impact autonomy has on reflectivity (path coefficients vary). This seems to be related to the different contexts, but all in all the impact is not large. The results show that work task analysis of dimensions of job design is worthwhile. The degree of complexity of the work has an effect on reflectivity and hence an indirect effect on self-organizational activities. The results suggest that the management could do more to support group processes and in the planning of production processes to support group initiative and self-organizational activities.

The goal of employees' work in a developmental and innovation process can be to generate ideas, but more often the goal is to create the right conditions so that new products or services can be produced. This puts demands on competence to be able to utilize all the discrepancies and uncertainties a change in product/service or work system implies as input to an iterative and reciprocal innovation process, and competence to proactively create conditions so that future changes can be dealt with efficiently. Knowledge-based organizations presuppose knowledge-based processes through the entire flow of the organization. The results show that group-based work organization can be a meaningful aspect to increase proactive behavior and coworker initiative to change and develop the organization.

There is good support for the assertion that transformational leadership, characterized by the leader "motivating employees to go beyond standard expectation by transforming followers' attitudes, beliefs and values" (Strauss et al., 2009, p. 282) increases proactivity (ibid). There is also good support for the assertion that supportive leadership benefits group processes (Kozlowski & Bell, 2004; Lantz & Laflamme, 1998) and the results of this study are in line with this. The results of the study also reveal that group processes mediate job design /leadership and proactive behavior. In other words, proactive behavior in work groups benefits from good group processes and the management has a meaningful task to support and facilitate interaction in the group and to improve the job design so that the group has a complete task. Reflectivity is a crucial component if individuals or groups are to reassess habitual work and social routines (West et al., 2004: Edmundson, 1999; Kozlowski & Bell, 2004). Good cooperation and an amicable social exchange lay the foundation for the group being able to challenge ingrained patterns and exchange thoughts on what could be done differently.

A critical approach to on the work conditions, the work itself and social interaction are promoted by a management that asks employees to reflect on the work itself and not just on the work results. West et al. (2004) claim that reflection is a key indicator for innovation and that leaders should "encourage reflexivity in teams – coach them to stop working". Supportive leadership and transformational leadership need not be incompatible with each other as long as support is not given to designing, controlling, and planning the group's work. Effective and innovative work groups need a considerable amount of autonomy, a challenging and complex work task, supportive leadership that both promotes group processes and a leadership that motivates the group to go beyond the stipulated task.

## Limitations

There is an inherent weakness in the survey instruments as a result of the risk of priming. There is a clear risk of the intuitive relationship between independent variables, such as autonomy, influencing the appraisal of dependent variables such as group initiative. Two methods were used (work task analysis and a survey) in order to reduce the risk of priming. The model showed better fit when based on work task analysis to capture job design, and job design captured by work task analysis has a larger impact on reflectivity than subjective appraisal of autonomy. The evidence points to primer-effects not influencing the relationship between autonomy and reflectivity. The concepts of group initiative and self-organizational activities are closely related. It is probable that there is a certain amount of consistency in the reply patterns. A confirmative factor analysis showed that these amounts were independent of each other, but the results of the factor analysis were acceptable and not optimal.

Merging data from different samples on the other hand provides a more substantial basis for the analysis which in turn gives more reliable results, but the results can also be affected by one of the included samples having a different reply pattern to the others. A number of analyses were carried out to check whether the included part-studies gave similar results and had similar patterns even if the effect magnitudes vary. All the tests for model fit gave good results.

## Future Research

Prior to any future research, it would be useful to examine the significance of the management giving support and also actively contributing to increasing the degree of reflectivity in the groups. In this study, it was not possible to examine this type of transformative leadership for the simple reason that we could not find organizations with sufficiently large numbers of managers who both gave support to work groups and had the ambition to coach them to "stop working - start thinking" as West et al. (2004) propose.

Inclusion criteria for most group research have been a) the group composition must be stable over time and b) the groups must have worked together for some time before the effects of group work are studied. But globalization has placed tough demands on productivity and efficiency and increasingly, in Europe and in other countries, work is being organized in a customary way in different types of line production. Are the group work models, several of which are based on McGrath's (1984) input-process-output model, applicable to loosely composed groups? Or project groups assembled for a specific task and then dissolved? It is a major task for future research to test whether models for how creativity, drive and innovation are created are applicable to loosely composed work groups.

## REFERENCES

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability. In K. J. Klein, & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Brav, A., Andersson, K., & Lantz, A. (2009). Group initiative and self-organizational activities in industrial work groups. *European Journal of Work and Organizational Psychology*, *3*, 347–377. doi:10.1080/13594320801960482.

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel Psychology*, *46*, 823–850. doi:10.1111/j.1744-6570.1993.tb01571.x.

Crant, J. M. (2000). Proactive behavior in organizations. *Journal of Management*, *3*, 435–462. doi:10.1177/014920630002600304.

Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*, 350–383. doi:10.2307/2666999.

Ehrelspiel, K., Kiewert, A., & Lindemann, U. (1999). *Integrierte produktentwicklung* [Integrated product development]. München, Germany: Hanser.

Fay, D., & Frese, M. (2001). The concept of personal initiative: An overview of validity studies. *Human Performance*, *14*, 97–124. doi:10.1207/S15327043HUP1401_06.

Frese, M., Garst, H., & Fay, D. (2007). Making things happen: Reciprocal relationships between work characteristics and personal initiative in a four-wave longitudinal structural equation model. *The Journal of Applied Psychology*, *4*, 1084–1102. doi:10.1037/0021-9010.92.4.1084 PMID:17638467.

Frese, M., Teng, E., & Wijnen, C. J. (1999). Helping to improve suggestion systems; predictors of making suggestions in companies. *Journal of Organizational Behavior*, *20*, 1139–1156. doi:10.1002/(SICI)1099-1379(199912)20:7<1139::AID-JOB946>3.0.CO;2-I.

Hacker, W. (2003). Action regulation theory: A practical tool for the design of modern work processes? *European Journal of Work and Organizational Psychology*, *12*, 105–130. doi:10.1080/13594320344000075.

House, J. S. (1981). *Work stress and social support*. Reading, MA: Addison Wesley.

James, L. R., Demaree, R. G., & Wolf, G. (1993). Rwg: An assessment of within-group interrater agreement. *The Journal of Applied Psychology*, *78*, 306–309. doi:10.1037/0021-9010.78.2.306.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: Structural equation models with SIMPLIS command language*. Chicago, IL: Scientific Software International.

Kim, S. (2009). Creativity and innovation: Imperatives for global business and development. *International Journal of Information Systems in the Service Sector*, *3*, 45–46. doi:10.4018/jisss.2009070103.

Kozlowski, S. W., & Bell, B. S. (2003). Work groups and teams in organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology* (pp. 333–375). Wiley. doi:10.1002/0471264385.wei1214.

Lagroue, H. J. (1998). The effectiveness of virtual facilitation for supporting group-decision making. *International Journal of Information and Decision Sciences*, *2*, 164–177.

Lantz, A., & Brav, A. (2007). Job design for learning in work groups. *Journal of Workplace Learning*, *5*, 269–285. doi:10.1108/13665620710757833.

Lantz, A., & Laflamme, L. (1996). Leadership, social support and work influence: A study of the group form of working in a Swedish psychiatric hospital. *Industrial Relations*, *4*, 693–724.

Lantz, A., & Laflamme, L. (1998). Exchange of social support, work influence and conditions for effective work. *The Journal of Occupational Health and Safety - Australia and New Zealand, 3*, 279–290.

Lantz, A., & Laflamme, L. (1998). Do differences in physicians' conceptions of work conditions and social support depend on gender or hierarchy? *The Journal of Occupational Health and Safety - Australia and New Zealand, 3,* 291 - 304.

Matsson, I.-B. (2001). *Reflektion och lärande i arbetsgrupper* [Reflection and learning in work groups]. Unpublished master's thesis, Mälardalens högskola, Eskilstuna, Sweden.

Medsker, G. J., Williams, L. J., & Holahan, P. J. (1994). *A review of current practices for evaluating causal McGrath, J. E. (1984). Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice Hall.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company*. Oxford, UK: Oxford University Press.

North, K., Friedrich, P., & Lantz, A. (2006). Selbstorganisation als Metakompetenz. [Self-organization as a metacompetence]. In J. Erpenbeck (Ed.), Metakompetenzen und Kompetenzentwicklung (pp. 137-208). Arbeitsgemeinschaft Betriebliche Weiterbildungsforschung. Berlin, Germany: ESM Satz und Grafik

North, K., & Güldenberg, S. (2008). *Produktive Wissensarbeit(er)* [Productive knowledge work(ers)]. Wiesbaden, Germany: Gabler. doi:10.1007/978-3-8349-8083-0.

Parker, S. K., Williams, H. M., & Turner, N. (2006). Modeling the antecedents of proactive behavior at work. *The Journal of Applied Psychology*, *91*, 636–652. doi:10.1037/0021-9010.91.3.636 PMID:16737360.

Pohlandt, A., Shulze, F., Debitz, U., Hänsgen, C., & Lüdecke, S. (2007). *Software with instruction manual: evaluation and design of task content in consideration of safety and health protection.* Germany: Bochum, InfoMediaVerlag e.k.

Richter, P., Hemman, E., & Pohlandt, A. (1999). Objective task analysis and the prediction of mental workload: Results of the application of an action-oriented software tool (REBA). In M. Wiethoff & F. R. H. Zijlstra (Eds.), New approaches for modern problems in work psychology (pp. 67-76). Series: WORC report 99.10.001. Tilburg, the Nederlands: University Press.

Sawyer, K. (2007). *The creative power of collaboration*. New York, NY: Basic books.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, N J: Earlbaum.

Skog, E. (2008). *Can four constructs of proactive behaviour merge into one?* Unpublished master's thesis, University of Uppsala, Sweden.

Strauss, K., Griffin, M., & Rafferty, A. (2009). Proactivity directed toward the team and organization: The role of leadership, commitment and role-breadth self-efficacy. *British Journal of Management*, *20*, 279–291. doi:10.1111/j.1467-8551.2008.00590.x.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Pearson.

Tschan, F. (2000). *Produktivität in kleingruppen: Was machen produktive Gruppen anders und besser?* [Productivity in small groups: What do productive groups do different and better?]. Bern, Switzerland: Huber.

Wagner, J. A. (1995). Studies of individualism-collectivism: Effects on cooperation in groups. *Academy of Management Journal*, *38*, 152–172. doi:10.2307/256731.

Wall, W. P. (2009). Developing global competitiveness in health-care: A Thai organization's perspective. *International Journal of Information Systems in the Service Sector*, *4*, 61–72. doi:10.4018/jisss.2009062905.

West, M. A., Hirst, G., Richter, A., & Shipton, H. (2004). Twelve steps to heaven: Successfully managing change through developing innovative teams. *European Journal of Work and Organizational Psychology*, *13*, 269–299. doi:10.1080/13594320444000092.

Woods, S. A., & West, M. A. (2010). *The psychology of work and organizations*. Handover, UK: Cengage Learning.

*Annika Lantz is an Associate Professor of Psychology at the Department of Psychology, Uppsala University, Sweden. She obtained her PhD and master's degree in Psychology from Stockholm University and is a licensed psychologist and psychotherapist. Her research interests include learning at work, job design, work organization, teamwork and management. She has initiated several larger European projects focusing on the effects of work organization for life-long learning, employment and innovation. At the department of Psychology, Uppsala University she is responsible for the research and training in work and organizational psychology and teaches courses within this domain. As a senior consultant at Fritz Change Company AB she is currently involved in a larger project within the Swedish school system concerning strategic management and the introduction of teamwork in schools. Her work is presented in several international journals and she has written books and book chapters covering a wide range of aspects of work and organization that promote employees' health, learning, proactive behavior and innovation processes within enterprises.*

# Implementing Business Processes:
## A Database Trigger Approach

*Wai Yin Mok, Department of Economics and Information Systems, University of Alabama in Huntsville, Huntsville, AL, USA*

*Charles F. Hickman, Department of Accounting and Finance, University of Alabama in Huntsville, Huntsville, AL, USA*

*Christopher D. Allport, Department of Accounting and Finance, University of Alabama in Huntsville, Huntsville, AL, USA*

## ABSTRACT

*Database triggers are database procedures that are executed automatically when certain events occur and conditions are met. This paper presents a design methodology that helps users implement business processes using database triggers. The contributions of this paper are as follows. First, the proposed methodology uses the Unified Modeling Language (UML). UML is a standard modeling language for the software industry and many commercial CASE (Computer-Aided Software Engineering) tools support UML. Second, many expensive ERP (Enterprise Resource Planning) software systems are employed to implement business processes. The methodology proposed by this paper produces triggers that can be executed on MySQL, an open-source database system that is free for download. Third, as an example of the usefulness of the proposed methodology, the authors present a case study making use of database triggers in a tax audit process. This process involves many steps that require human intervention, and thus is typical of business processes.*

*Keywords:* *Business Processes, Computer-Aided Software Engineering (CASE), Database Triggers, Enterprise Resource Planning (ERP) Software, MySQL, Unified Modeling Language (UML)*

## INTRODUCTION

Due to the technological advances in the last twenty to thirty years, large corporations have come to realize that Information Technology (IT) is becoming a primary driving force in doing business (Chang, 2005). Automated business processes are guiding various stakeholders in making decisions and are able to instantly notify stakeholders of progress. There are many ERP software systems on the market today. Examples are Systems Analysis and Program Development (SAP), Oracle and PeopleSoft. Despite the benefits provided by ERP software, the business models supported by this software may not coincide with the culture of the adopting institutions. Many horror stories attest that either the culture of the institution has to change

DOI: 10.4018/ijkbo.2013040103

dramatically, or the software has to be modified significantly (Madara, 2007). Either way the result is undesirable.

As an alternative to expensive ERP software, this paper presents a design methodology that, if followed, will reduce the cost of developing a work flow system for a business process. The proposed methodology takes advantages of free open-source software. Since most ERP software cost tens of thousands of dollars, the proposed methodology provides substantial savings.

At the center of the proposed methodology are active database systems, which serve as platforms on which workflow systems are built. To understand their importance in our approach, note that a typical business process encounters a lot of human-generated events. These events are external to the resulting workflow system. Many of these external events occur simultaneously and any single one of them might lead to numerous modifications to the backend database. Therefore, if the platform on which the workflow system is built is able to respond to external events and act accordingly, most of the programming responsibilities can be delegated to the platform itself. As a result, a lot of programming effort can be saved. As they are designed to react to events and act accordingly, active database systems are therefore well-suited to be such supporting platforms.

To gain an understanding of active database systems, note that database systems in the 70s and 80s were passive, meaning that they did not act until instructed. Today's database systems are different. Most of them support triggers. Consequently, they are able to react to external events; and as a result they have become active. Triggers, a relatively new addition to database systems, are named database procedures that are automatically executed when certain events occur and conditions are met. As such, triggers are able to monitor any changes to the business environment and modify the backend database accordingly to satisfy established business rules and logic. Example systems that support triggers include Oracle, IBM DB2, and MySQL. For this paper, MySQL is our choice because (1) it is free, (2) it supports triggers, and other procedures and functions, and (3) it is commonly used as a backend database system in many web applications.

Our methodology has several stages. Given a business process *P*, we first construct a model for *P* by using an activity diagram in UML. UML is a standard modeling language in the software industry, and therefore has a well-understood syntax. Therefore, many software CASE tools on the market support UML. Hence, building our methodology upon UML will make it easier to integrate into existing CASE tools. After the model is complete, we then map it to a MySQL database, which contains a collection of tables and triggers. The triggers of the database will then implement the business logic of *P*. Lastly, we build a web application that lets the agents enter information into the database. The information will be further processed by the triggers and the MySQL database will provide feedback to the agents for their use in carrying out the activities in *P*.

This paper is organized as follows. The second section introduces activity diagrams and triggers. In the third section, we discuss how we map activity diagrams to MySQL tables and triggers. We study the property of termination for a collection of triggers in the fourth section . The case study for this paper is presented in the fifth section. We make concluding remarks in the sixth section.

## FUNDAMENTALS

### Activity Diagrams

The main tool in the UML for modeling business processes is activity diagrams (Booch, Jacobson, & Rumbaugh, 2005). Activity diagrams are similar to traditional flowcharts, but with additional modeling constructs. Like flowcharts, activity diagrams are able to show branches of control. However, activity diagrams are also able to show concurrency, which is not allowed in traditional flowcharts. Since concurrency is an important concept of business processes, activity diagrams are superior in this regard.

Activity diagrams have many modeling constructs. However, not all of them are relevant for this paper. Some relevant modeling constructs of activity diagrams are activities, control flows, object values, decisions, forking, joining and swimlanes (Booch et al., 2005). Each of these constructs is explained as follows:

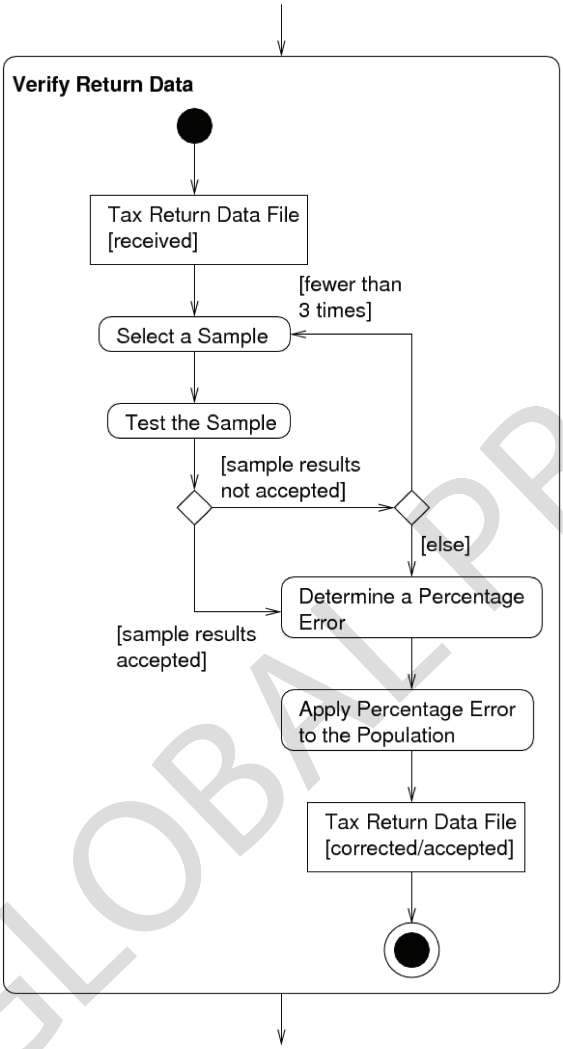
- **Activities:** There are two kinds of activities: basic and nested. Basic activities are atomic, meaning that their details are not revealed. In addition, a basic activity either executes entirely or not at all. On the other hand, a nested activity is defined by an activity diagram, meaning that zooming into a nested activity will discover another activity diagram. As such, the execution of a nested activity ultimately expands into the execution of individual basic activities. An activity, either basic or nested, may take some time to complete. When the context is clear, a basic activity or a nested activity is simply referred to as an activity to avoid being too verbose.
- **Control Flows:** When an activity completes, flow of control passes to the next activity. This control flow is specified by using a flow arrow to show the path of control.
- **Object Values:** During the execution of an activity diagram, objects may change states. We specify the objects that are involved in an activity diagram by placing these objects in the diagram, connected by control flows to the activities that produce or consume them. We represent the state of an object by naming its state in brackets below the object's name.
- **Decisions:** Sequential control flows are common. However, oftentimes different paths of control in a business process are necessary. To determine the correct path, decisions must be made. Each outgoing flow of a decision is guarded by a boolean expression, which is evaluated on entering the decision. The boolean expressions of the outgoing flows should not overlap and they should cover all possibilities. This means one and only one outgoing flow will be chosen at a decision.

**Example 1:** The defining activity diagram of a nested activity called "verify return data" in a tax audit process is shown in Figure 1. Notationally, we use a round-cornered rectangle, called an activity node, to represent an activity in an activity diagram. The top filled circle in Figure 1 represents the initiation of this activity diagram. The bottom filled circle, with an encircled circle, represents its completion. Each of the four basic activities in Figure1 is represented by an activity node. Diamonds, on the other hand, represent decisions. In our example, there are two decision nodes and their outgoing control flows are guarded by boolean expressions. One decision node tests if a sample of receipts is acceptable. If so, based on the sample receipts, we determine the percentage error that will be projected onto the entire population. If not, one or more additional samples of receipts will be collected. If all samples are considered unacceptable, we will have to negotiate with the company to determine an agreeable percentage error. Lastly, the two rectangles in the activity diagram represent the different states of the object "tax return data file." Initially, it is in the "received" state. Finally, it is in the "corrected" or "accepted" state, depending on whether we have to negotiate with the company on an agreeable percentage or not.

In addition to supporting branching, activity diagrams also support forking and joining, which are used in modeling concurrency. These terms are explained as follows:

- **Forking:** Forking is used to initiate concurrency and joining to synchronize two or more concurrent flows of control. At a fork, a single flow of control splits into two or more concurrent flows of control.
- **Joining:** At a join, multiple concurrent flows synchronize, meaning that each waits until all incoming flows have reached the join, at which point one flow of control continues on.

*Figure 1. A nested activity that verifies tax return data*

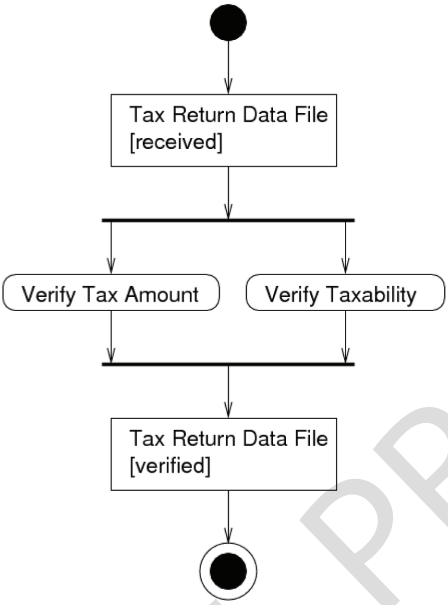

Verify Return Data
Tax Return Data File [received]
[fewer than 3 times]
Select a Sample
Test the Sample
[sample results not accepted]
[else]
Determine a Percentage Error
[sample results accepted]
Apply Percentage Error to the Population
Tax Return Data File [corrected/accepted]

In UML, we use a synchronization bar to specify the forking to and joining of multiple flows of control. A synchronization bar is rendered as a thick horizontal or vertical line.

**Example 2:** The top thick horizontal line in Figure 2 represents a synchronization bar, which specifies the forking of two parallel flows of control from a single one. These two flows of control lead to the two activities in the figure. After both activities complete, they lead to another synchronization bar, represented by the bottom thick horizontal line. The bottom synchronization bar specifies the joining of two parallel flows of control into a single one.

- **Swimlanes:** Eventually, the activities in an activity diagram must be carried out by some people or objects. For this purpose, we partition the activity nodes in an activ-

*Figure 2. Two concurrent activities*



ity diagram into groups. In the UML, each group is called a swimlane because each group is divided from its neighbors by vertical lines. The activities in each swimlane eventually will be carried out by one or more parties, which can be real-world or software objects.

**Example 3:** In Figure 3, there are two swimlanes, one for the team leader and one for a team member of an auditing team. By using swimlanes, we can clearly see the responsible party of an activity node in the activity diagram.

## Database Triggers

Many database systems, commercial or open-source, support triggers. A trigger is a named database procedure that is associated with a table and is automatically executed when a particular event occurs for the table. In most database systems, the events that activate a trigger are inserting a row into, deleting a row from, and updating a row in the table with which the trigger is associated. In most cases, the timing of activating a trigger can come before or after the event. As a result, six different types of triggers are BEFORE-INSERT, AFTER-INSERT, BEFORE-UPDATE, AFTER-UPDATE, BEFORE-DELETE and AFTER-DELETE.

**Example 4:** Figure 4 shows a create-table statement for a table called “tax return data file” in MySQL syntax. It also shows the definition of an AFTER-INSERT trigger associated with the table. The table has three columns. The columns “case number” and “time stamp” are self-explanatory. The column “status” records the state of a tax return in the database. Two allowable states of a tax return are “received” and “verified,” as enumerated in the create-table statement. The AFTER-INSERT trigger “tax return data file after insert” is defined on the table “tax return data file.” The trigger will be activated after a row is inserted into the table. According to the MySQL manual, the OLD and NEW keywords enable us to access columns in the rows affected

*Figure 3. A nested activity that includes swimlanes*

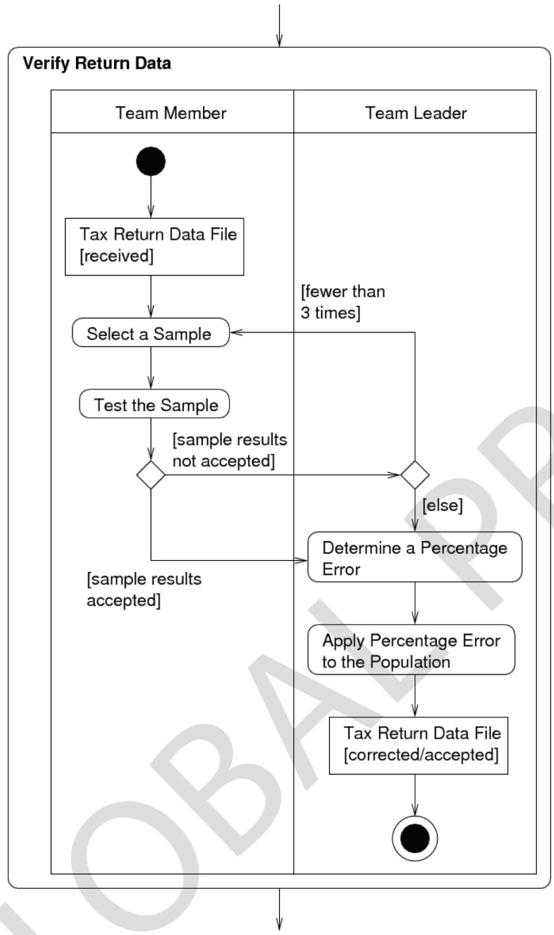


*Figure 4. A MySQL table and an associated AFTER-INSERT trigger*

```
create table tax_return_data_file (
  case_number smallint unsigned not null,
  status      enum('received',
                   'verified') not null,
  time_stamp  datetime not null
);

create trigger tax_return_data_file_after_insert after insert on
tax_return_data_file for each row begin
  if new.status = 'received' then
    insert into verify_tax_amount(case_number, status, time_stamp,
    agent) values
      (new.case_number, 'start', sysdate(), 'John Smith');
    insert into verify_taxability(case_number, status, time_stamp, agent)
    values
      (new.case_number, 'start', sysdate(), 'Sue Jones');
  end if;
end//
```

by a trigger. In an INSERT trigger, only NEW.col_name can be used; there is no old row. In a DELETE trigger, only OLD. col_name can be used; there is no new row. In an UPDATE trigger, we can use OLD. col_name to refer to the columns of the row being updated before it is updated and NEW.col_name to refer to the columns of the row being updated after it is updated. Overall, the AFTER-INSERT trigger in Figure 4 states that when a new tax return with the status "received" is inserted into the table, the trigger will insert a row into the table "verify_tax_amount" and a row into the table "verify_taxability," signaling the beginning of these two activities. Note that a new tax return must have the status "received" in order to activate the trigger. If a new tax return with the status "verified" is inserted into the table, nothing will happen. .

The action of a trigger may activate other triggers. For example, the insert statements of the AFTER-INSERT triggerinFigure4 may activate other triggers. In fact, it is possible that triggers may activate one another forever if they are not designed correctly. In the fourth section, we shall discuss the property of termination for a collection of triggers.

## MAPPING ACTIVITY DIAGRAMS TO MYSQL TABLES AND TRIGGERS

We now turn to the details on mapping an activity diagram to MySQL tables and triggers.

**Rule 1:** For each start node, each end node, each synchronization bar that represents joining, and for each activity in an activity diagram, we create a table with the same name. This table records the state of the activity at a particular time, indicated by the time stamp. The allowable states are "immediately_done," "start," "complete_without_errors," "complete_with_errors," and "abort." The state "immediately_done" means an activity is immediately finished once it has started. This also means the activity does not take any time to complete. The state "start" means an activity has just begun and will take some time to finish. The other three states indicate how an activity ends. It may end successfully, meaning that the activity will enter the "complete_without_errors" state. It may end unsuccessfully, meaning that the activity will enter the "complete_with_errors" state. If an abnormal event happens during the execution of the activity so that the activity does not complete at all, the activity will enter the "abort" state. The time the activity enters any of these states is recorded in the "time_stamp" column. If the activity enters the state "complete_with_errors" or "abort," we may want to enter an error message in the column "error" to record the event that causes the errors.

**Example 5:** We generate the create-table statements in Figure 5 for the activity diagram in Figure 2. All of the tables created in Figure 5 have the same table structure. Therefore, only the create-table statement for the start node is shown in full detail. The activity nodes "start" and "end" do not require any time to finish. Therefore, they will only enter the state "immediately_done." For the other activity nodes "verify_tax_amount," "verify_taxability," and the synchronization bar that represents joining, they all need time to finish. Thus, the states that they may enter are "start," "complete_without_errors," "complete_with_errors," and "abort." .

**Rule 2:** For each synchronization bar that represents forking in an activity diagram, the forking logic is handled by the trigger associated with the table for the activity node or object node that comes before the synchronization bar in the activity diagram. For each synchronization bar that represents joining, we create a table. Its associated trigger contains the logic of synchronizing the different flows of

*Figure 5. The resulting MySQL tables of applying Rule 1 to Figure 2*

```
create table start_node (
  case_number smallint unsigned not null,
  status            enum('immediately_done',
                         'start',
                         'complete_without_errors',
                         'complete_with_errors',
                         'abort') not null,
  time_stamp      datetime not null,
  agent           varchar(50) default null,
  error           varchar(50) default null
);

create table verify_tax_amount (
          .
          .
          .
);

create table verify_taxability (
          .
          .
          .
);

create table synchronization_bar_join (
          .
          .
          .
);

create table end_node (
          .
          .
          .
);
```

control. The idea is to check if all of the incoming activities have completed without any error. If so, then the synchronization succeeds; otherwise, it fails.

**Example 6:** There are two synchronization bars in Figure 2. The top one represents forking and the bottom one joining. To differentiate them, we call the top one "synchronization_bar_fork" and the bottom one "synchronization_bar_join."

Recall that there is not table created for the synchronization bar that represents forking. Instead, the forking logic is handled by the trigger associated with the table for the activity node or object node that comes before the synchronization bar in the activity diagram. In Figure 2, the object node "tax_return_data_file" comes before "synchronization_bar_fork," and thus the forking logic is handled in the trigger for the table "tax return data file," which is originally shown in Figure 4 and is repeated in Figure 6 for convenience. To do so, the AFTER-INSERT trigger in Figure 6 inserts a row into the table "verify_tax_amount" and a row into the table "verify_taxability," signaling the beginning of the two activities.

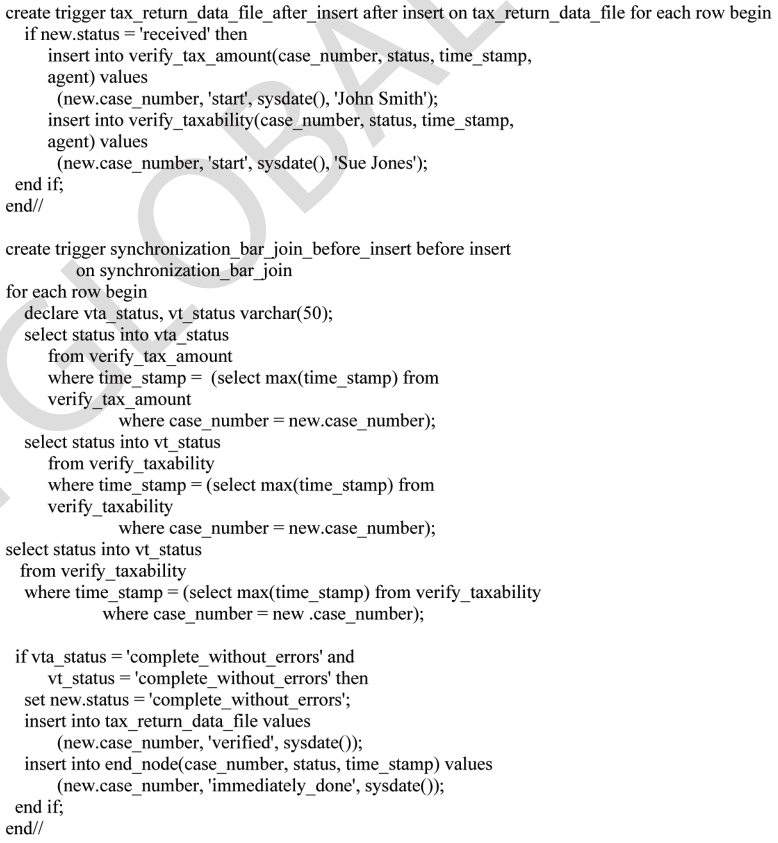
We now explain the BEFORE-INSERT trigger "synchronization_bar_join_before_insert" in Figure 6, whose purpose is to synchronize the activities "verify_tax_amount' and "verify_taxability." When the activity "verify_tax_amount" completes successfully, the agent who carried out the activity will enter a row into the "verify_tax_amount" table to signal the completion of the activity. This will activate the AFTER-INSERT trigger "verify_tax_amount_after_insert" in Figure 7, which will insert a row into the table "synchronization_bar_join." In turn, it will activate the trigger "synchronization_bar_join_before_insert" in Figure 6. The same thing happens when the activity "verify_taxability" completes successfully. When the trigger "synchronization_bar_join_before_insert" is activated, it first finds the current states of these activities. If both of them are in the state "complete_without_errors," then the trigger inserts a row into the table "tax_return_data_file" with the status "verified." At the same time, it inserts a row into the table for the end node, which signals the completion of the entire workflow.

**Rule 3:** For each decision node in an activity diagram, the logic of the decision is handled in the trigger associated with the table of the node comes before the decision node in the activity diagram.

Our naming convention for the table of an internal activity of a nested activity is that, for each internal activity, the nested activity's name becomes the prefix of the name of the internal activity's associated table. As an example, we can see in Figure 8 that the table with the name "verify_return_data_test_the_sample" has two parts: the name of the nested activity of which the internal activity is a part and the name of the internal activity.

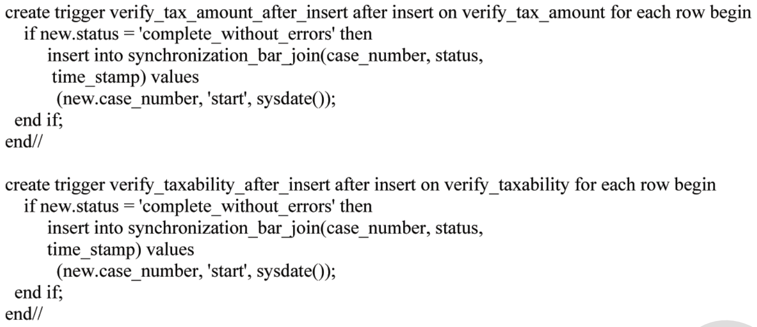
**Example 7:** There are two decision nodes in Figure 1. The activity that comes before them is "test the sample." Hence, as shown in Figure 8, the logic for both of these decision nodes is handled in the trigger for the table "verify return data test the sample."

*Figure 6. Triggers for forking and joining*



```
create trigger tax_return_data_file_after_insert after insert on tax_return_data_file for each row begin
   if new.status = 'received' then
        insert into verify_tax_amount(case_number, status, time_stamp,
        agent) values
          (new.case_number, 'start', sysdate(), 'John Smith');
        insert into verify_taxability(case_number, status, time_stamp,
        agent) values
          (new.case_number, 'start', sysdate(), 'Sue Jones');
  end if;
end//

create trigger synchronization_bar_join_before_insert before insert
            on synchronization_bar_join
for each row begin
   declare vta_status, vt_status varchar(50);
   select status into vta_status
        from verify_tax_amount
        where time_stamp =  (select max(time_stamp) from
        verify_tax_amount
                    where case_number = new.case_number);
   select status into vt_status
        from verify_taxability
        where time_stamp = (select max(time_stamp) from
        verify_taxability
                    where case_number = new.case_number);
select status into vt_status
  from verify_taxability
   where time_stamp = (select max(time_stamp) from verify_taxability
                  where case_number = new .case_number);

  if vta_status = 'complete_without_errors' and
        vt_status = 'complete_without_errors' then
   set new.status = 'complete_without_errors';
   insert into tax_return_data_file values
         (new.case_number, 'verified', sysdate());
   insert into end_node(case_number, status, time_stamp) values
         (new.case_number, 'immediately_done', sysdate());
  end if;
end//
```

*Figure 7. Another AFTER-INSERT trigger*

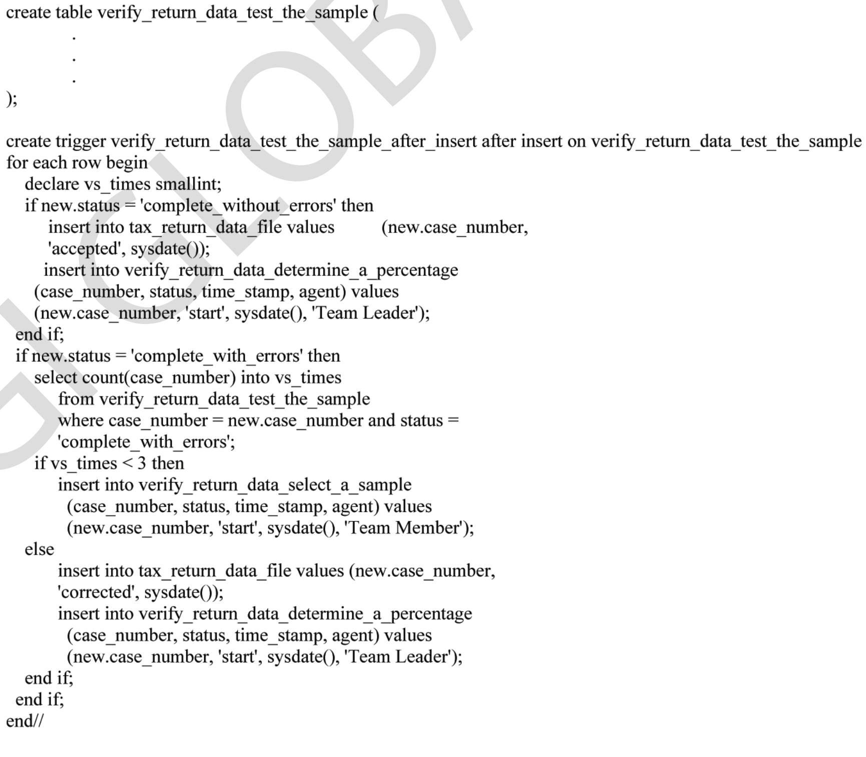


```
create trigger verify_tax_amount_after_insert after insert on verify_tax_amount for each row begin
   if new.status = 'complete_without_errors' then
        insert into synchronization_bar_join(case_number, status,
         time_stamp) values
          (new.case_number, 'start', sysdate());
 end if;
end//

create trigger verify_taxability_after_insert after insert on verify_taxability for each row begin
   if new.status = 'complete_without_errors' then
        insert into synchronization_bar_join(case_number, status,
        time_stamp) values
          (new.case_number, 'start', sysdate());
 end if;
end//
```

## TERMINATION OF TRIGGER FIRING

Since triggers may fire one another and the execution might go on forever, an important property of a set of triggers is termination (Baralis & Widom, 2000). Analysis techniques for termination of a set of triggers are abundant (Aiken, Hellerstein, & Widom, 1995; Baralis & Widom, 2000; Montesi, Bertino, & Bagnato, 2003). Yet we shall show that even though these techniques are successful to a certain degree, the termination problem of a set of triggers in general cannot be solved algorithmically. Note

*Figure 8. An AFTER-INSERT trigger that models a decision node*



```
create table verify_return_data_test_the_sample (
            .
            .
            .
);

create trigger verify_return_data_test_the_sample_after_insert after insert on verify_return_data_test_the_sample
for each row begin
    declare vs_times smallint;
    if new.status = 'complete_without_errors' then
        insert into tax_return_data_file values        (new.case_number,
        'accepted', sysdate());
       insert into verify_return_data_determine_a_percentage
     (case_number, status, time_stamp, agent) values
     (new.case_number, 'start', sysdate(), 'Team Leader');
  end if;
  if new.status = 'complete_with_errors' then
     select count(case_number) into vs_times
         from verify_return_data_test_the_sample
         where case_number = new.case_number and status =
         'complete_with_errors';
     if vs_times < 3 then
         insert into verify_return_data_select_a_sample
           (case_number, status, time_stamp, agent) values
           (new.case_number, 'start', sysdate(), 'Team Member');
    else
         insert into tax_return_data_file values (new.case_number,
         'corrected', sysdate());
         insert into verify_return_data_determine_a_percentage
           (case_number, status, time_stamp, agent) values
           (new.case_number, 'start', sysdate(), 'Team Leader');
    end if;
  end if;
end//
```

that the same termination problem was studied in Bailey, Dong, and Ramamohanarao (1998) in the context of logic programming. Nevertheless, our proof is far simpler and straightforward.

In Mok, Palvia, and Paper (2006), we made use of abacus, a theoretical computer, to prove an undecidability result of state charts. In this paper, we use abacus again to prove that the termination of a set of triggers is undecidable. To make this paper self-contained, we reproduce the definition of abacus from Mok et al. (2006).

An abacus is a theoretical computer in the sense that it has an unlimited number of registers and each register can store a number of any size (Boolos & Jeffrey, 1989). No real computers have these two properties since a real computer has a fixed number of memory cells and each memory cell can only store a number up to a certain number of digits. Since we are studying the theoretical aspects of triggers, it is not inappropriate to ignore these physical limitations. Registers in an abacus are like elements in an array of a programming language. Thus, we use a similar notation *reg*(*i*) to denote the *i*th register in an abacus, where $i \geq 1$. Since other kinds of objects such as strings and negative integers that can be manipulated by computers can be encoded and decoded as nonnegative integers (Sipser, 1997), in this paper we only consider nonnegative integers. Consequently, numbers stored in registers are integers greater than or equal to zero.

An abacus can add one to a register and subtract one from a register if the current number in the register is greater than zero. Since no other operations are possible with an abacus, (in this sense) it is a very primitive computer. Graphically, these additions and subtractions are represented as nodes, as shown in Figure 9. The intended operation of a node is written as its label. Because negative numbers are not allowed, there are only two possibilities with subtractions: either the number in a register is already zero or is greater than zero. In the first case, the intended subtraction is ignored and the outgoing arrow marked with *e* is followed to locate the next operation.[1] In the second case, one is subtracted from the register and the outgoing arrow not marked with e is followed to find the next operation. Note that any outgoing arrows in Figure 9 can be absent. For example, if the outgoing arrow of a node of addition is missing, no more operations will be executed after the addition is done. Similarly, either one or both of the two outgoing arrows of a node of subtraction can be missing. An abacus program is a directed graph of these nodes and arrows. One of the nodes in a program is specified as the start node, which means its operation is the first one to be executed.

Following the notation in Montesi et al. (2003), a trigger $r_i$ is written as follows:

$r_i$: $E[C] \rightarrow A_1 ; ...; A_n$.

Trigger $r_i$ means that when event *E* occurs and condition *C* is true, then the actions $A_1$, ..., $A_n$ are carried out in the order specified. As mentioned in Widom and Ceri (1996), event *E*, condition *C*, and actions $A_1$, ..., $A_n$ are all optional.

**Theorem 1:** The termination of a set of triggers is undecidable.

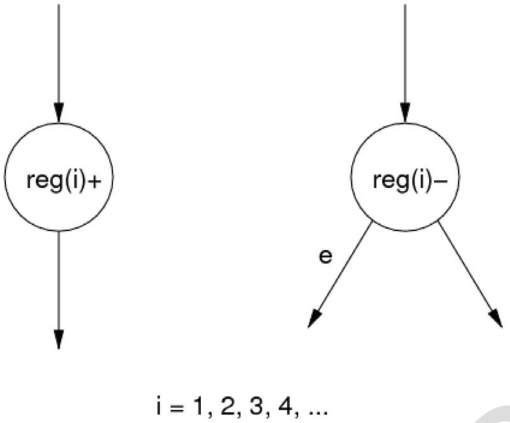
- **Proof:** Given an abacus program *P*, we derive algorithmically a set of triggers that computes the same set of functions. The resulting triggers have access to the same infinite array of registers. Suppose *P* has n nodes $N_1$, $N_2$, ..., $N_n$. We define *n* events $E_1$, $E_2$, ..., $E_n$ for the purpose that when $E_i$ is raised, we execute the operation of $N_i$. Next we define n triggers $r_1$, $r_2$, ..., $r_n$ where $r_i$ corresponds to $N_i$, $1 \leq i \leq n$. If $N_i$ is labeled with *reg*(*j*) + and has an out going arrow that points at node $N_k$, then $r_i$ is defined as:

$r_i$: $E_i \rightarrow reg(j) := reg(j) + 1$; raise $E_k$ .

Similarly, if $N_i$ is labeled with *reg*(*j*)+ and does not have any outgoing arrow, then $r_i$ is simply defined as:

$r_i$: $E_i \rightarrow reg(j) := reg(j) + 1$.

*Figure 9. Two elementary operations of an abacus*



For a node of subtraction, there are four cases to consider based on whether the outgoing arrow marked with e is missing or the outgoing arrow not marked with *e* is missing. Since these four cases are very similar, we just consider two of them: both outgoing arrows are absent and both are present. In the first case, if $N_i$ is labeled with *reg*(*j*) -, $r_i$ is defined as:

$r_i$: $E_i$ → if *reg*(*j*)> 0 then *reg*(*j*):= *reg*(*j*) – 1;
end if

In the second case, if $N_i$ is labeled with *reg*(*j*)- and the outgoing arrow marked with *e* points at node $N_k$ and the outgoing arrow not marked with *e* points at node $N_l$, $r_i$ is defined as:

$r_i$: $E_i$ → if *reg*(*j*) > 0 then *reg*(*j*):= *reg*(*j*) – 1;
raise $E_l$; else raise $E_k$; end if

Thus, each trigger fires up the other trigger by raising its corresponding event. This transformation is algorithmic and therefore the proof is complete.

Although in general there is not any algorithmic solution to determine if a set of triggers will terminate execution, there is a simple solution if the set of triggers is obtained by the mapping rules of this paper. This result is proved as the following lemma.

**Lemma 1:** For an activity diagram *P*, let *S* be the resulting set of triggers obtained by Rules 1, 2, and 3 in this paper. If each activity in *P* can only execute a certain number of times, the set *S* will terminate.

- **Proof:** Since the number of activities in *P* is fixed, if the set *S* executes forever, then at least one activity in *P* is executing an unlimited number of times. However, if each activity in *P* can only execute a certain number of times, then *P* will terminate and thus *S* will terminate as well. .

**Example 8:** There is a loop in the activity diagram in Figure 1. Thus, potentially this activity diagram may execute forever. However, the activity "select a sample" can at most execute three times. Thus, the loop cannot go on forever. Therefore, the activity diagram will terminate and the resulting set of triggers also terminates as well.

# CASE STUDY

## Overview

Sales tax audits are conducted to measure taxpayer compliance with the sales tax laws. Two of the primary objectives of these audits are (1) to generate revenue for the state and (2) to foster voluntary compliance with sales tax laws. The current Case Study analyzes the benefits of database triggers in this context in two ways. First, given that the work load on tax auditors and managers in the taxing authority can be considerable, the management and control of audit workflow through the use of activity diagrams and swimlanes can be considerable. Second, the tax audit process can be simplified with the use of before and after triggers when performing specific activities in the audit process. We specifically consider, "Testing the Sample," and show that the benefits of this technology can be considerable.

## Managing Tax Audit Workflow

When considering the role database triggers can play in a sales tax audit, we first consider the importance of activity diagrams and swimlanes presented earlier in the paper. These capabilities are important for tax auditors and the taxing authority because several audits maybe in process for any one auditor and clearly many more will be in the inventory of the team leaders and managers. For this reason controlling and managing the completion of audits is an important responsibility.

As discussed previously in the paper, the tax audit process includes multiple steps. These steps can be considered as the required audit workflow:

1. Initiate a new audit case.
2. Select a sample of tax data.
3. Test the sales tax data for the sample.
4. Determine the percentage of error in the sample sales tax data.
5. Apply the percentage of error for the sample to the population data.
6. Close the audit case.

To demonstrate the feasibility of using MySQL triggers to manage tax audit workflow, we created several web-based forms written in php that serve as an interface to the MySQL database and the triggers presented earlier. The main form is shown in Figure 10, which allows adding a new case, checking the status of every case and updating the status of a particular case.

When the audit team member receives the name of the taxpayer that will be audited, the taxpayer will first be inserted into the database as a new case through a form shown in Figure 11. The audit team member also enters a case number and the beginning status of the case, which usually is "start" that signals the initiation of the tax audit process. Clicking on the link "check all cases" on the main form will show the status of every case in the database.

After adding the new case, the activity diagram moves the audit process through to the next step, selecting a sample. By checking the process of all audit cases, a team leader or manager could verify that case1had moved to step two in the audit process. Figure 12 shows that Case 1 started at 15:51:13, and the second step, "select_a_sample", started at 15:51:16.

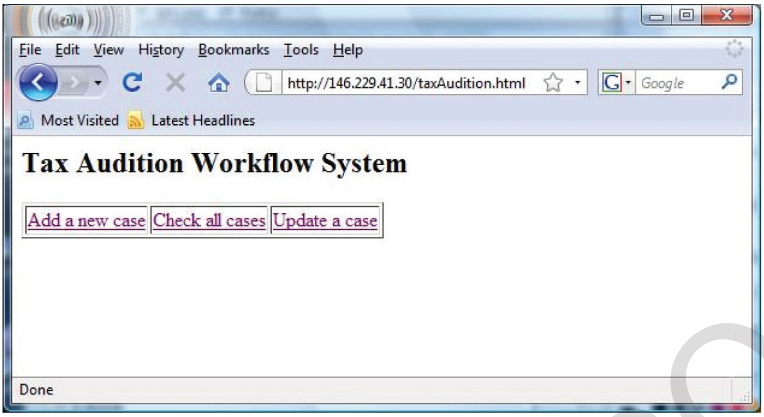
When the activity "select_a_sample" is complete, the status of the case needs to be updated. The auditor will return to the main form of Figure 10 and click on the link "update_a_case." The next form in Figure 13 allows the agent to select the desired case to update and Figure 14 displays the form with the current activity of that case. This form further allows the auditor to enter the new status of the activity and the agent who is responsible for the update.

In this manner, the MySQL database and the triggers point to the next activity to be preformed, and thus it guides the tax audit team in this tax audit process. It also records the time stamps of the events and the person who is responsible for each activity in the process. Assuming the process does not encounter any anomaly, Figure 15 displays the final status of Case 1, and the final status of the tax return data file.

## Automating the Tax Audit Process

Managing the workflow for a sales tax audit is a major benefit of database triggers; however, additional tools available via these triggers can also be of benefit within the specific steps of a tax audit. To demonstrate this idea, we use the third step in the sales tax audit, "Test_the_sales_tax_data_for_the_sample" (also see

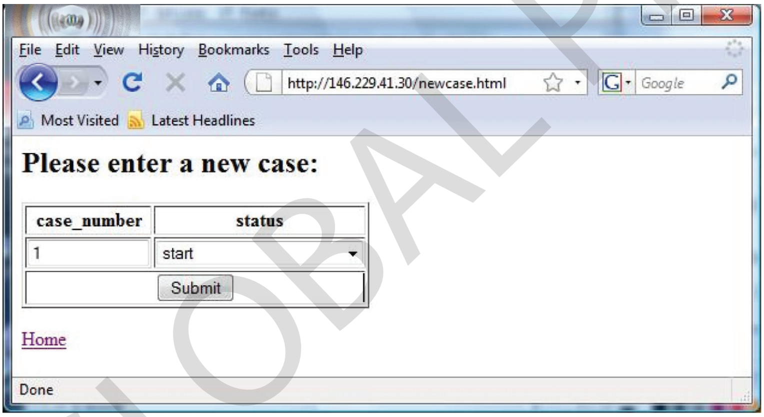
*Figure 10. Tax audit initiation*


File Edit View History Bookmarks Tools Help
http://146.229.41.30/taxAudition.html
Google
Most Visited
Latest Headlines
Done

## Tax Audition Workflow System

| Add a new case | Check all cases | Update a case |
|---|---|---|

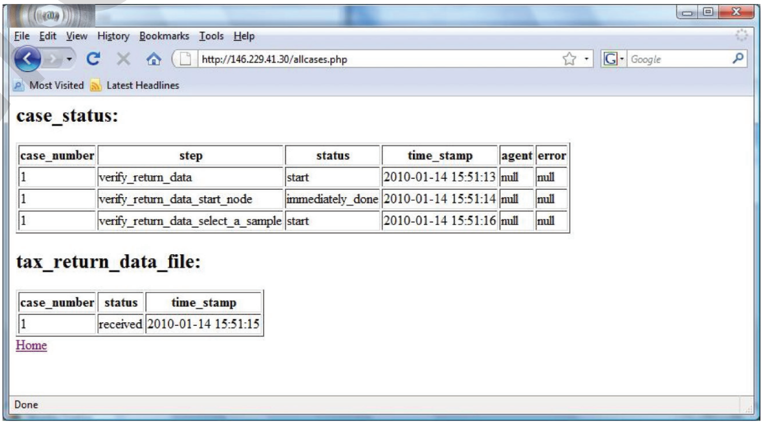
*Figure 11. Initiating a new case*


File Edit View History Bookmarks Tools Help
http://146.229.41.30/newcase.html
Google
Most Visited
Latest Headlines
Submit
Home
Done

## Please enter a new case:

| case_number | status |
|---|---|
| 1 | start |

*Figure 12. Selecting a sample*


File Edit View History Bookmarks Tools Help
http://146.229.41.30/allcases.php
Google
Most Visited
Latest Headlines
Home
Done

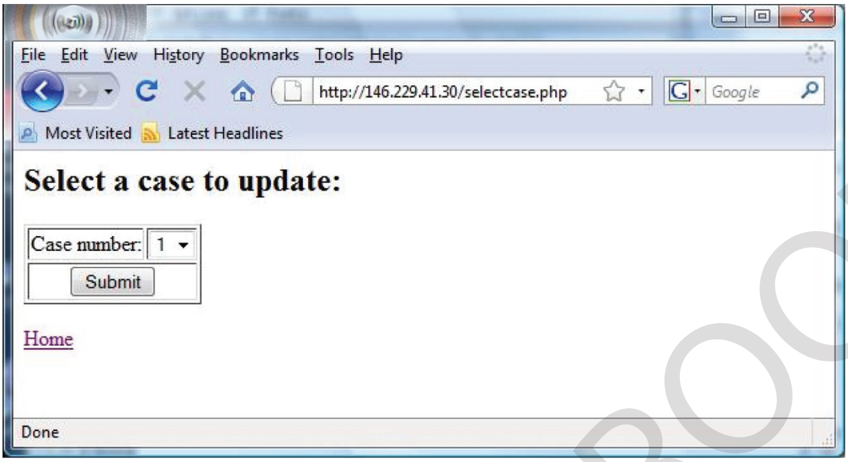
### case_status:

| case_number | step | status | time_stamp | agent | error |
|---|---|---|---|---|---|
| 1 | verify_return_data | start | 2010-01-14 15:51:13 | null | null |
| 1 | verify_return_data_start_node | immediately_done | 2010-01-14 15:51:14 | null | null |
| 1 | verify_return_data_select_a_sample | start | 2010-01-14 15:51:16 | null | null |

### tax_return_data_file:

| case_number | status | time_stamp |
|---|---|---|
| 1 | received | 2010-01-14 15:51:15 |

*Figure 13. Case workflow update*


File Edit View History Bookmarks Tools Help
http://146.229.41.30/selectcase.php
Google
Most Visited
Latest Headlines
Done

**Select a case to update:**

| | |
|---|---|
| Case number: | 1 |
| Submit | |

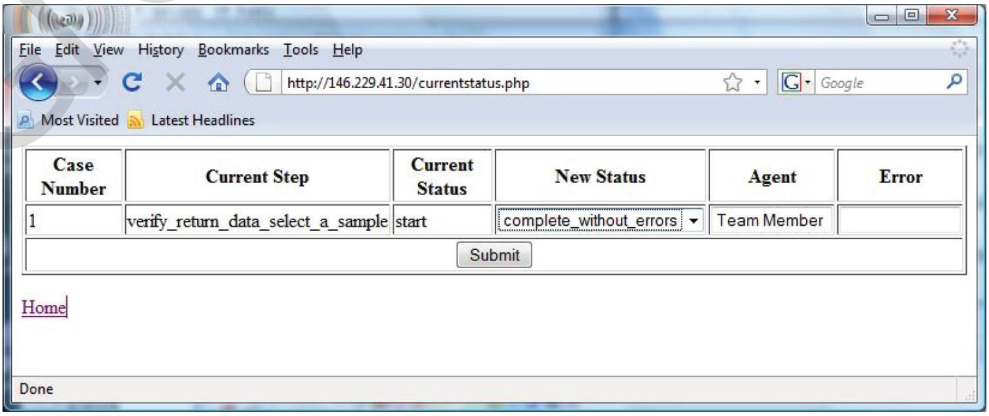
Home

verify_return_data_test_the_sample in Figure 15) in our case study as an example. This step requires the auditor to consider two different issues. Specifically, the auditor will consider each of the sample tax transactions and test (1) whether the transaction is taxable and (2) whether the amount remitted to the tax authority for the transaction is mathematically accurate.

We designed two tables and two triggers, shown in Figure 16, to address these issues.

The sample tax transactions are stored in the table "transactions" and the results of the calculation are stored in the table "percentageErrorTable." Two triggers, one of which is BEFORE-INSERT and the other AFTER-INSERT, are defined on the table "transactions." Whenever a new sample transaction is inserted into the table "transactions," the BEFORE-INSERT trigger is first activated. It checks whether the zip code of the transaction is an Alabama zip code, which is in between of 35004 and36925. If so, then it sets the columns al_trans and al_sales_tax to the values of trans_amt and sales_tax respectively. If not, then the columns al_trans and al_sales_tax will take on the default null values. After the execution of the BEFORE-INSERT trigger, the AFTER-INSERT trigger is activated. It sums up the columns al_trans

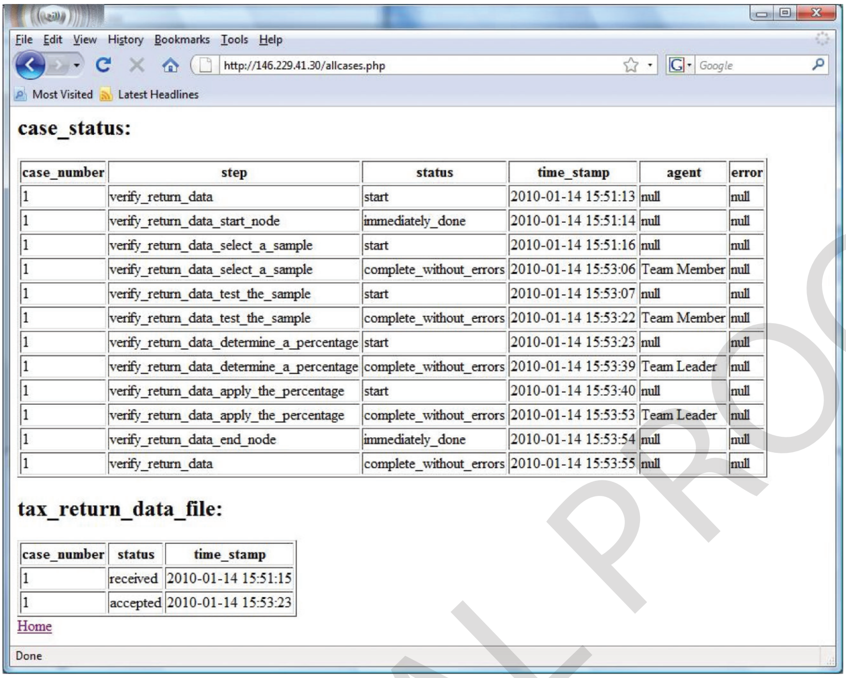
*Figure 14. Completing sample selection*


File Edit View History Bookmarks Tools Help
http://146.229.41.30/currentstatus.php
Google
Most Visited
Latest Headlines
Done

| Case Number | Current Step | Current Status | New Status | Agent | Error |
|---|---|---|---|---|---|
| 1 | verify_return_data_select_a_sample | start | complete_without_errors | Team Member | |
| Submit | | | | | |

Home

*Figure 15. Completed tax audit workflow*


File Edit View History Bookmarks Tools Help
http://146.229.41.30/allcases.php
Google
Most Visited
Latest Headlines

## case_status:

| case_number | step | status | time_stamp | agent | error |
|---|---|---|---|---|---|
| 1 | verify_return_data | start | 2010-01-14 15:51:13 | null | null |
| 1 | verify_return_data_start_node | immediately_done | 2010-01-14 15:51:14 | null | null |
| 1 | verify_return_data_select_a_sample | start | 2010-01-14 15:51:16 | null | null |
| 1 | verify_return_data_select_a_sample | complete_without_errors | 2010-01-14 15:53:06 | Team Member | null |
| 1 | verify_return_data_test_the_sample | start | 2010-01-14 15:53:07 | null | null |
| 1 | verify_return_data_test_the_sample | complete_without_errors | 2010-01-14 15:53:22 | Team Member | null |
| 1 | verify_return_data_determine_a_percentage | start | 2010-01-14 15:53:23 | null | null |
| 1 | verify_return_data_determine_a_percentage | complete_without_errors | 2010-01-14 15:53:39 | Team Leader | null |
| 1 | verify_return_data_apply_the_percentage | start | 2010-01-14 15:53:40 | null | null |
| 1 | verify_return_data_apply_the_percentage | complete_without_errors | 2010-01-14 15:53:53 | Team Leader | null |
| 1 | verify_return_data_end_node | immediately_done | 2010-01-14 15:53:54 | null | null |
| 1 | verify_return_data | complete_without_errors | 2010-01-14 15:53:55 | null | null |

## tax_return_data_file:

| case_number | status | time_stamp |
|---|---|---|
| 1 | received | 2010-01-14 15:51:15 |
| 1 | accepted | 2010-01-14 15:53:23 |

Home

Done

and al_sales_tax, calculates the correct sales tax amount, and percentage error. Finally, it updates the sole row in the table "percentageErrorTable" to record these results.

To interface with these tables, we also designed several web-based forms. They are shown in Figures 17, 18, and 19. Figures 17 and 18 are self-explanatory.

Figure 19, on the other hand, displays all sample transactions and the results of the calculation. Note that for those transactions that do not have Alabama zip codes, the rows have null values under the al_trans and al_sales_tax columns.

## CONCLUSIONS

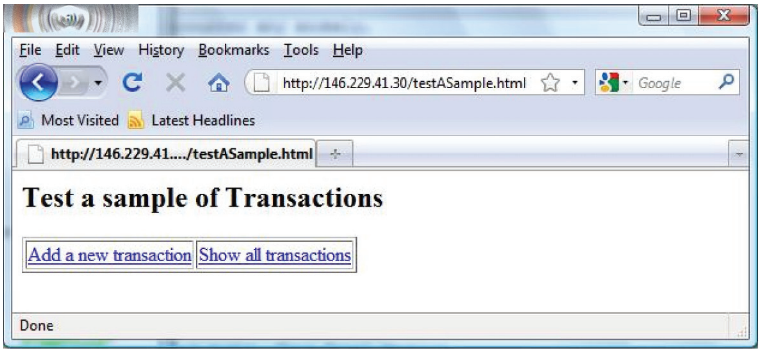
This paper presents a design methodology that helps users implement database triggers for business processes. We first model the business process using an activity diagram in the UML. We then map the model to a MySQL database, allowing the triggers of the database to implement the logic of the activity diagram. Finally, we provide an example of the database triggers in a tax audit case study. Our design methodology provides two important advantages for practice. First, our design methodology utilizes the UML, which is the standard modeling language for the software industry and is therefore applicable for many business users. Second, our methodology utilizes MySQL, which is a free, open-source database system. This option could provide considerable cost savings when compared to well known ERP systems such as SAP, Oracle, and PeopleSoft.

Our next step is to automate the mapping rules of this paper. That is, given an activity diagram, we will develop a program that automatically generates the required triggers and tables. Although it may seem that the number of required tables can be large, it is bounded by a constant factor from the number of activities. As a result, the number of triggers and tables required for an activity diagram is linear with respect to the number of activities in the diagram.

*Figure 16. Tables and triggers for testing a sample of transactions*

```
create table transactions (
  trans_no      smallint unsigned not null,
  cust_id       smallint unsigned not null,
  zip integer   unsigned not null,
  trans_amt     decimal(12,4) not null,
  sales_tax     decimal(12,4) not null,
  al_trans      decimal(12,4) default null,
  al_sales_tax  decimal(12,4) default null
);

create table percentageErrorTable (
  total_sample_trans_amt      decimal(12,4) default null,
  total_sample_tax_collected  decimal(12,4) default null,
  correct_tax_amt             decimal(12,4) default null,
  percentage_error            decimal(12,4) default null
);

insert into percentageErrorTable values (null, null, null, null);

create trigger transactions_before_insert before insert on transactions
for each row begin
  if 35004 <= new.zip and new.zip <= 36925 then
    set new.al_trans = new.trans_amt;
    set new.al_sales_tax = new.sales_tax;
  end if;
end//

create trigger transactions_after_insert after insert on transactions
for each row begin
  declare sampleALTrans  decimal(12,4);
  declare sampleALTax    decimal(12,4);
  declare correctALTax   decimal(12,4);
  declare error          decimal(12,4);

select sum(al_trans) into sampleALTrans
  from transactions;
select sum(al_sales_tax) into sampleALTax
  from transactions;
set correctALTax = sampleALTrans * 0.08;
set error = (correctALTax - sampleALTax) / sampleALTax;
update percentageErrorTable
  set total_sample_trans_amt = sampleALTrans, total_sample_tax_collected = sampleALTax,
      correct_tax_amt = correctALTax, percentage_error = error;
end//
```

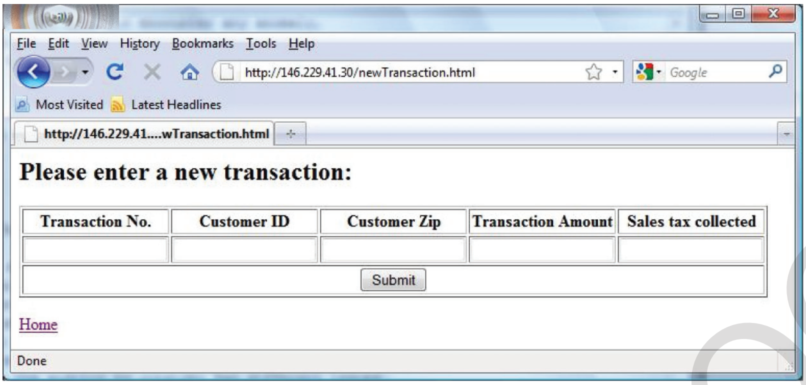
*Figure 17. Test a sample of transactions*


File Edit View History Bookmarks Tools Help
http://146.229.41.30/testASample.html
Google
Most Visited
Latest Headlines
http://146.229.41..../testASample.html
Done

Test a sample of Transactions

Add a new transaction | Show all transactions

Figure 18. Enter a new transaction


File Edit View History Bookmarks Tools Help
http://146.229.41.30/newTransaction.html
Google
Most Visited
Latest Headlines
http://146.229.41....wTransaction.html
Submit
Home
Done

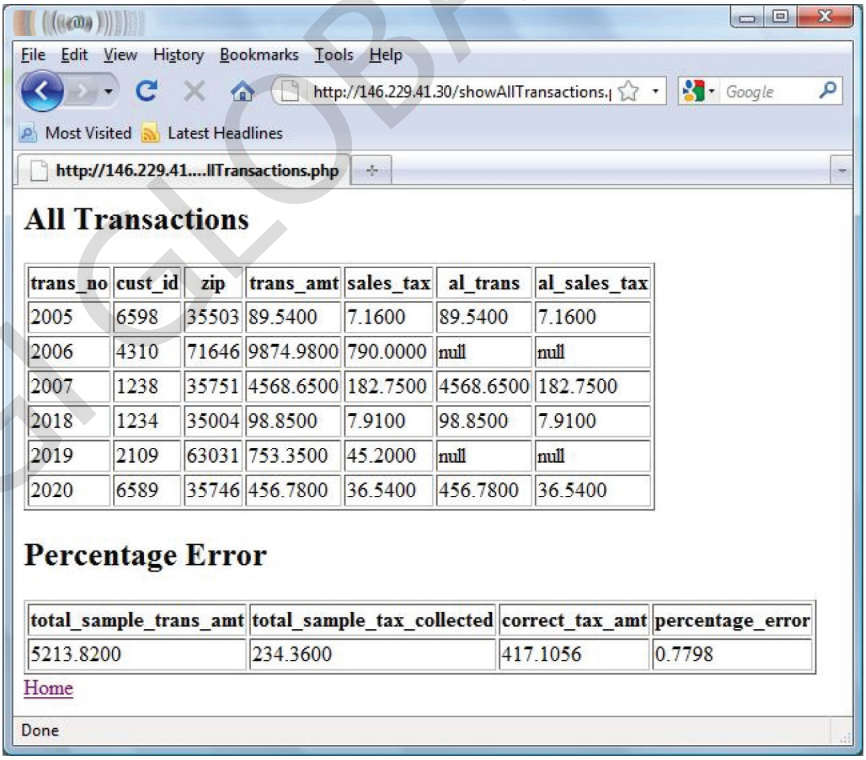
Please enter a new transaction:

| Transaction No. | Customer ID | Customer Zip | Transaction Amount | Sales tax collected |
|---|---|---|---|---|
| | | | | |

Consequently, the program will have relatively fast run-time.

Finally, we stress that we propose an inexpensive alternative to commercial ERP systems, which may easily cost thousands of dollars. However, it is quite obvious that commercial ERP systems provide more functionality than the proposed approach of this paper. So, the proposed approach is a good way for a company to experiment with an ERP system.

Figure 19. Display all transactions and percentage error


File Edit View History Bookmarks Tools Help
http://146.229.41.30/showAllTransactions.
Google
Most Visited
Latest Headlines
http://146.229.41....llTransactions.php
Home
Done

All Transactions

| trans_no | cust_id | zip | trans_amt | sales_tax | al_trans | al_sales_tax |
|---|---|---|---|---|---|---|
| 2005 | 6598 | 35503 | 89.5400 | 7.1600 | 89.5400 | 7.1600 |
| 2006 | 4310 | 71646 | 9874.9800 | 790.0000 | null | null |
| 2007 | 1238 | 35751 | 4568.6500 | 182.7500 | 4568.6500 | 182.7500 |
| 2018 | 1234 | 35004 | 98.8500 | 7.9100 | 98.8500 | 7.9100 |
| 2019 | 2109 | 63031 | 753.3500 | 45.2000 | null | null |
| 2020 | 6589 | 35746 | 456.7800 | 36.5400 | 456.7800 | 36.5400 |

Percentage Error

| total_sample_trans_amt | total_sample_tax_collected | correct_tax_amt | percentage_error |
|---|---|---|---|
| 5213.8200 | 234.3600 | 417.1056 | 0.7798 |

After the company gains more experience and desires more functionality, it may consider a more expensive alternative.

## REFERENCES

Aiken, A., Hellerstein, J. M., & Widom, J. (1995). Static analysis techniques for predicting the behavior of active database rules. *ACM Transactions on Database Systems*, *20*, 3–41. doi:10.1145/202106.202107.

Bailey, J., Dong, G., & Ramamohanarao, K. (1998). Decidability and undecidability results for the termination problem of active database rules. *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Seattle, WA (pp. 264–273).

Baralis, E., & Widom, J. (2000). An algebraic approach to static analysis of active database rules. *ACM Transactions on Database Systems*, *25*, 269–332. doi:10.1145/363951.363954.

Booch, G., Jacobson, I., & Rumbaugh, J. (2005). *The unified modeling language user guide* (2nd ed.). Pearson.

Boolos, G. S., & Jeffrey, R. C. (1989). *Computability and logic* (3rd ed.). Cambridge University Press.

Chang, J. F. (2005). *Business process management systems: Strategy and implementation*. CRC Press. doi:10.1201/9781420031362.

Madara, E. (2007). A recipe and ingredients for ERP failure. *Articles base: Free online articles directory*. Retrieved from www.articlesbase.com/software-articles/

Mok, W. Y., Palvia, P., & Paper, D. (2006). On the computability of agent-based work-flows. *Decision Support Systems*, *42*, 1239–1253. doi:10.1016/j.dss.2005.10.010.

Montesi, D., Bertino, E., & Bagnato, M. (2003). Refined rules termination analysis through transactions. *Information Systems*, *28*, 435–456. doi:10.1016/S0306-4379(02)00025-X.

Sipser, M. (1997). *Introduction to the theory of computation*. Boston, MA: PWS Publishing Company.

J. Widom, & S. Ceri (Eds.). (1996). *Active database systems: Triggers and rules for advanced database processing*. Morgan Kaufmann.

## ENDNOTES

[1] *e* means empty

*Wai Yin Mok received the BS, MS, and PhD degrees in computer science from Brigham Young University in 1990, 1992, and 1996, respectively. He is an associate professor of information systems at the University of Alabama in Huntsville. His papers have appeared in the* ACM Transactions on Database Systems, IEEE Transactions on Knowledge and Data Engineering, Data & Knowledge Engineering, Decision Support Systems, Information Processing Letters, *and the* Journal of Database Management.

*Charles F. Hickman is a Clinical Assistant Professor of Taxation and Business Legal Studies at the University of Alabama in Huntsville. He has a Bachelor of Science Degree in General Business, a Juris Doctor and a Master of Letters and Laws in Taxation. Prior to beginning his educational career he had a twenty year career in the private practice of law, as an attorney for the Illinois Department of Revenue and in public accounting. He is licensed as an attorney and CPA in the State of Alabama.*

*Christopher D. Allport is an Assistant Professor of Accounting at the University of Alabama in Huntsville. He has a PhD in Business Administration from Virginia Tech. He has published in* Information Systems Research and Information Systems Journal*, and his general research interests include various aspects of information systems and business judgments and decision making.*

# A Framework for Synthesizing Arbitrary Boolean Queries Induced by Frequent Itemsets

*Animesh Adhikari, Department of Computer Science, S. P. Chowgule College, Margao, Goa, India*

## ABSTRACT

*Frequent itemsets determine the major characteristics of a transactional database. It is important to mine arbitrary Boolean queries induced by frequent itemsets. In this paper, the author proposes a simple and elegant framework for synthesizing arbitrary Boolean queries using conditional patterns in a database. Both real and synthetic databases were used to evaluate the experimental results. The author presents an algorithm for mining a set of specific itemsets in a database and a model of synthesizing a query in a database. Finally, the author discusses an application of the proposed framework for reducing query processing time.*

*Keywords:* *Boolean Query, Conditional Pattern, Data Mining, Generator of a Boolean Expression, Synthesis of Query*

## INTRODUCTION

An itemset could be thought as the basic type of pattern in a transactional database. Itemset patterns influence research in knowledge discovery in databases (KDD) in the following ways: Firstly, many interesting algorithms have been reported on mining itemset patterns in a database (Agrawal & Srikant, 1994; Han et al., 2000; Savasere al., 1995). Secondly, many patterns are defined based on the itemset patterns in a database. They may be called as derived patterns. Some examples of derived patterns are positive association rules (Agrawal et al., 1993), negative association rules (Antonie & Zaïane, 2004), and conditional patterns (Adhikari & Rao, 2008) in a database. A good amount of work has been reported on mining / synthesizing such derived patterns. Thirdly, the solutions of many problems are based on the analysis of patterns in a database. Such applications (Adhikari & Rao, 2008; Wu et al., 2005) process patterns in a database for the purpose of making some decisions. Thus, the mining and analysis of itemset patterns in a database is an interesting as well as important issue. Also, mining Boolean expressions induced by frequent itemsets could lead to significant nuggets of knowledge with many potential applications in market basket data analysis, web usage mining, social network analysis and bioinformatics. There are two important goals of this paper. First, we design a framework for synthesizing an arbitrary Boolean expression

DOI: 10.4018/ijkbo.2013040104

induced by frequent itemsets. Afterwards, we present a technique for reducing query processing time using such synthesized knowledge.

The *support* (Agrawal et al., 1993) of an itemset $X$ in database $D$ could be defined as the fraction of transactions in $D$ containing all the items of $X$, denoted by $S(X, D)$. The importance of an itemset could be judged by its support. An itemset $X$ is *frequent* in $D$ if $S(X, D) \geq$ *minimum support* ($\alpha$). Let $SFIS(D)$ be the set of frequent itemsets in $D$. Frequent itemsets determine the major characteristics of a database. Wu et al. (2005) have proposed a solution of inverse frequent itemset mining. Authors argued that one could efficiently generate a synthetic market basket dataset from the frequent itemsets and their supports. Let $X$ and $Y$ be two itemsets in $D$. The characteristics of $D$ are revealed more by the pair ($X$, $S(X, D)$) than that of ($Y$, $S(Y, D)$), if $S(X, D) > S(Y, D)$. So it is important to study frequent itemsets more than infrequent itemsets. Hence, we propose a framework for synthesizing arbitrary Boolean queries induced by frequent itemsets in $D$. Zhao et al. (2006) have proposed the BLOSOM framework for mining arbitrary Boolean expressions. The framework suffers from the following limitations:

- It does not handle NOT operator effectively.
- Let $\{a, b, c\}$ be a frequent itemset of our interest. We wish to mine some functions induced by $\{a, b, c\}$. It proposes a framework to mine minimal generators of (i) closed OR-clauses, (ii) closed AND-clauses, (iii) closed maximal min-DNF, and (iv) closed maximal min-CNF. It requires establishing a mapping from the space of minimal generators to the space of arbitrary Boolean expressions, so that we could study the desired Boolean expressions induced by $\{a, b, c\}$. Thus, the BLOSOM framework might not provide the knowledge of Boolean expression that one wishes to study.
- A specific framework for a specific type of Boolean expressions is introduced.

Moreover, most of the existing works (Pei & Han, 2000; Bonchi & Lucchese, 2005) have attempted to answer queries during the mining process. First, we propose here a simple and elegant approach for synthesizing arbitrary Boolean queries induced by frequent itemsets. The proposed framework of synthesizing Boolean queries is based on conditional patterns in $D$. Afterwards, we have presented an application of such synthesized knowledge by presenting a framework for answering arbitrary queries.

First we explain the concept of conditional pattern and then we present a framework for synthesizing Boolean queries induced by frequent itemsets in $D$. The concept of conditional pattern is not new (Adhikari & Rao, 2008). For the purpose of completeness, we discuss the notion of conditional pattern in this section. Let $X = \{a, b, c\}$ be an itemset in $D$. The study of items in $X$ might be incomplete if we have only the following information about $X$: (i) the supports of $X$ and its subsets, (ii) the association rules generated from $X$. The answers to some queries on the items of $X$ are not immediately available from (i) and (ii). A few examples of such queries are given below:

- Find the support that a transaction contains item $a$, but not items $b$ and $c$ with respect to frequent itemset $\{a, b, c\}$.
- Find the support that a transaction contains items $a$ and $b$ but not the item $c$ with respect to frequent itemset $\{a, b, c\}$.

The above queries correspond to a specific type of pattern in a database. Some of these patterns could have significant supports, since $\{a, b, c\}$ is a frequent itemset. In general, let $X$ be a frequent itemset in $D$. If we wish to study the distribution of items of itemset $Y \subseteq X$, but not the items of itemset $X$-$Y$, then such analysis of items is not be immediately available. Analysis of such patterns is interesting, since their supports could be high. Therefore, one needs to identify this type of patterns. Let $\langle Y, X \rangle$ be a pattern that a transaction in a database contains all the items of itemset $Y$, but not the items of itemset $X$-$Y$, for a given itemset $X$ in $D$. Let $S\langle Y, X, D \rangle$ be the support that a transaction in $D$ contains all the items of $Y$, but not the items of $X$-$Y$, for a given itemset $X$ in $D$. The pattern of type $\langle Y, X \rangle$ is called *conditional pattern*. A conditional pattern has two components viz., *pattern itemset* ($Y$) and *reference itemset* ($X$). Thus, a conditional pattern $\langle Y, X\rangle$ is associated with two values viz., $S\langle Y, X, D\rangle$ and $S(X, D)$. $S\langle Y, X, D\rangle$ and $S(X, D)$ are called *conditional support* (*csupp*) and *reference support* (*rsupp*) of conditional pattern $\langle Y, X\rangle$ in $D$, respectively. A conditional pattern is interesting if the conditional support is greater than or equal to the *minimum conditional support* ($\delta$) and the reference support is greater than or equal to $\alpha$. $\alpha$ and $\delta$ are user-defined inputs to a conditional pattern mining algorithm. For the proposed problem, one needs to mine all the conditional patterns irrespective of their conditional supports.

Let us come back to above queries. We present the following figures to understand these queries better.

The shaded region of Figure 1(i) contains the set of transactions containing the item $a$, but not the items $b$ and $c$ with respect to itemset $\{a, b, c\}$ in $D$. Thus, $S\langle \{a\}, \{a, b, c\}, D\rangle = S(\{a\}, D) - S(\{a, b\}, D) - S(\{a, c\}, D) + S(\{a, b, c\}, D)$. The shaded region of Figure 1(ii) contains the set of transactions containing the items $a$ and $b$, but not the item $c$ with respect to itemset $\{a, b, c\}$ in $D$. Thus, $S\langle \{a, b\}, \{a, b, c\}, D\rangle = S(\{a, b\}, D) - S(\{a, b, c\}, D)$. A conditional pattern $\langle Y, X\rangle$ in a database is *trivial* if $Y = X$. A trivial conditional pattern is known when the corresponding frequent itemset gets extracted from the database. Thus, the trivial conditional patterns get mined during the mining of frequent itemsets. In the following example, we identify the conditional patterns with reference to an itemset in the database.

**Example 1:** Let $D$ be a database containing the following transactions: $\{a, b\}$, $\{a, b, c, d\}$, $\{a, b, c, h\}$, $\{a, b, g\}$, $\{a, b, h\}$, $\{a, c\}$, $\{a, c, d\}$, $\{b\}$, $\{b, c, d, h\}$, $\{b, d, g\}$. Let $\alpha$ be 0.2. The conditional patterns with reference to $X = \{a, b, c\}$ in $D$ are given in Table 1.

All the conditional patterns mined with reference to a given frequent itemset are not interesting. The interesting conditional patterns with reference to itemset $\{a, b, c\}$ in $D$ are given in Table 2.

We observe that *csupp* $\langle Y, X, D\rangle \leq$ *supp* $(Y, D)$, but still *csupp* $\langle Y, X, D\rangle$ could be high, for an itemset $Y \subset X$. Thus, it is necessary to study such patterns in a database for an effective analysis of items in the itemset.

The pattern itemset of a conditional pattern with reference to itemset $X = \{a_1, a_2, \ldots, a_m\}$ is of the form $b_1 \wedge b_2 \wedge \ldots \wedge b_m$, where $b_i = a_i$, or $\neg a_i$, for all $i = 1, 2, \ldots, m$. Let $\psi(X)$ be the set of all such pattern itemsets with reference to $X$ excluding the pattern itemset $\neg a_1 \wedge \neg a_2 \wedge \ldots \wedge \neg a_m$. $\psi(X)$ is called a *generator* of the Boolean expressions induced by $X$. $\psi(X)$ contains $2^m$-1 pattern itemsets. A pattern itemset of the corresponding conditional pattern is also called a *minterm*, or *standard product*. We will see later, how every Boolean expression of items of $X$ could be constructed using pattern itemsets in $\psi(X)$. In particular, let $X = \{a, b, c\}$. Then, $\psi(X) = \{a\wedge b\wedge c, a\wedge b\wedge\neg c, a\wedge\neg b\wedge c, a\wedge\neg b\wedge\neg c, \neg a\wedge b\wedge c, \neg a\wedge b\wedge\neg c, \neg a\wedge\neg b\wedge c\}$. The Boolean expression $\neg b\wedge c$ could be re-written as $(a\wedge\neg b\wedge c) \vee (\neg a\wedge\neg b\wedge c)$. Every Boolean expression can be expressed as a sum of some pattern

*Figure 1. Shaded regions in (i) and (ii) correspond to conditional supports of $\langle\{a\}, \{a, b, c\}\rangle$ and $\langle\{a, b\}, \{a, b, c\}\rangle$ in D, respectively*

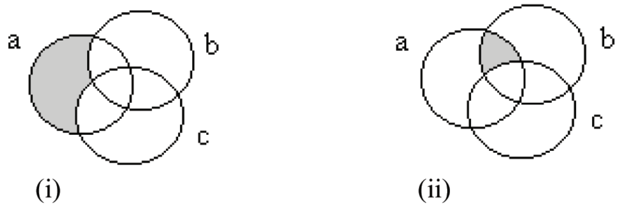


*Table 1. Conditional patterns with respect to {a, b, c} in D*

| conditional patterns | csupp | conditional patterns | csupp |
|---|---|---|---|
| $\langle\{a\}, X\rangle$ | 0 | $\langle\{a, c\}, X\rangle$ | 0.2 |
| $\langle\{b\}, X\rangle$ | 0.2 | $\langle\{b, c\}, X\rangle$ | 0.1 |
| $\langle\{c\}, X\rangle$ | 0 | $\langle X, X\rangle$ | 1.0 |
| $\langle\{a, b\}, X\rangle$ | 0.3 | | |

itemsets in the generator. A Boolean expression expressed as a sum of pattern itemsets is said to be in *canonical form*. Each pattern itemset corresponds to a set of transactions in *D*. In the following, we show how each pattern itemset with reference to $\{a, b, c\}$ corresponds to a set of transactions in *D*.

The shaded region in Figure 2(i) contains the set of transactions containing the items *a*, *b* and *c* with respect to $\{a, b, c\}$. Thus, it corresponds to the pattern itemset of $\langle\{a, b, c\}, \{a, b, c\}\rangle$. The shaded region in Figure 2(ii) contains the set of transactions containing the items *a*, and *b*, but not the item *c* with respect to $\{a, b, c\}$. Thus, it corresponds to the pattern itemset of $\langle\{a, b\}, \{a, b, c\}\rangle$. The shaded region in Figure 2(iii) contains the set of transactions containing the items *a* and *c*, but not the item *b* with respect to $\{a, b, c\}$. Thus, it corresponds to the pattern itemset of $\langle\{a, c\}, \{a, b, c\}\rangle$. The shaded region in Figure 2(iv) contains the set of transactions containing the item *a*, but not the items *b* and *c* with respect to $\{a, b, c\}$. Thus, it corresponds to the pattern itemset of $\langle\{a\}, \{a, b, c\}\rangle$. The shaded region in Figure 3(i) contains the set of transactions containing the items *b* and *c*, but not the item *a* with respect to $\{a, b, c\}$. Thus, it corresponds to the pattern itemset of $\langle\{b, c\}, \{a, b, c\}\rangle$. The shaded region in Figure 3(ii) contains the set of transactions containing the item *b*, but not the items *a* and *c* with respect to $\{a, b, c\}$. Thus, it corresponds to the pattern itemset of $\langle\{b\}, \{a, b, c\}\rangle$. Finally, the shaded region in Figure 3(iii) contains the set of transactions containing the item *c*, but not the items *a* and *b* with respect to $\{a, b, c\}$. Thus, it corresponds to the pattern itemset of $\langle\{c\}, \{a, b, c\}\rangle$.

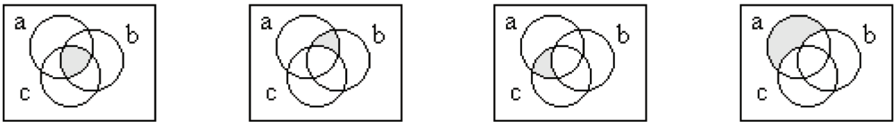
The data in a transaction is binary in nature. One could construct $2^{2^m}$ Boolean expressions (functions) for *m* Boolean items (variables). Thus, the number of Boolean expressions induced by all frequent itemsets in a database could be large. But, there are only $2^m$-1 conditional patterns corresponding to a frequent itemset of size *m*. Thus, it could be better to mine the generator of an itemset and synthesize the desired Boolean expressions afterwards.

The rest of the paper is organized as follows. We discuss related work in the second section. We present some results in the third section. In the fourth section, we propose an algorithm for synthesizing generators. The results of the experiments are presented in the fifth section. In the sixth section, we discuss how one could reduce the average query processing time by applying synthesized knowledge, and then we conclude this paper in the seventh section.

*Table 2. Non-trivial conditional patterns with respect to {a, b, c} at δ = 0.2 and α = 0.2*

| conditional pattern | *csupp* | *rsupp* |
|---|---|---|
| $\langle\{b\}, \{a, b, c\}\rangle$ | 0.2 | 0.2 |
| $\langle\{a, b\}, \{a, b, c\}\rangle$ | 0.3 | 0.2 |
| $\langle\{a, c\}, \{a, b, c\}\rangle$ | 0.2 | 0.2 |

*Figure 2. Generator of {a, b, c}*



(i) $a \wedge b \wedge c$ (ii) $a \wedge b \wedge \neg c$ (iii) $a \wedge \neg b \wedge c$ (iv) $a \wedge \neg b \wedge \neg c$

## RELATED WORK

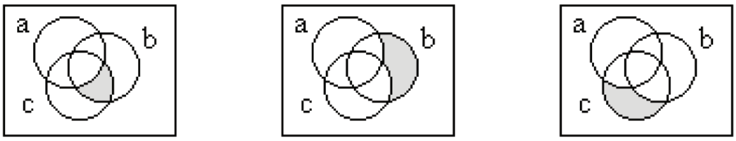
Mining conjunctive Boolean expressions (Agrawal & Srikant, 1994; Han et al., 2000; Savasere et al., 1995) have been well studied within the context of frequent itemset mining. Several implementations of mining conjunctive Boolean expressions have been reported recently (FIMI, 2004). The maximum-entropy approach to support estimation of a general Boolean expression is proposed by Pavlov et al. (2000). Itemset mining typically results in large amount of redundant itemsets. Several approaches such as closed itemsets, non-derivable itemsets and generators have been suggested for reducing the amount of itemsets losslessly. Muhonen and Toivonen (2006) have proposed a pruning method based on combining techniques for closed and non-derivable itemsets that allows further reductions of itemsets. This reduction is done without loss of information, that is, the complete collection of frequent itemsets can still be derived from the collection of closed non-derivable itemsets. Shima et al. (2004) have proposed a technique of mining closed and minimal monotone disjunctive normal forms. The proposed technique for synthesizing Boolean expression is somewhat different from the technique proposed by Shima et al. Within the association rule context, there has been previous work on mining negative rules (Wu et al., 2004).

Query processing and query optimization have been studied well for relational database management systems (Yu & Meng, 1997). One of the main aims of query optimization is to reduce the number of I/O operations.

Pei and Han (2000) have studied various convertible constraints and monotone constraint on frequent pattern growth mining. A significant amount of research on efficient processing of frequent itemset queries has been done in recent years, focusing mainly on constraint handling and reusing results of previous queries (Baralis & Psaila, 1999; Cheung et al., 1996; Meo, 2003; Morzy et al., 2000). But the proposed framework is based on post-processing of mining results.

Mediavilla and Lee (2009) have presented the idea of using statistical methods to model federated data marts. Once data marts are modeled, reduced sets of distributed data can be imported and used to approximately reconstruct a federated data mart. Approximate queries can then be obtained from the reconstructed federated data mart. Sun and Li (2009) have proposed usage control models to access XML databases and compared with an authorization model designed for object-oriented databases.

*Figure 3. Generator of {a, b, c}*



(i) $\neg a \wedge b \wedge c$ (ii) $\neg a \wedge b \wedge \neg c$ (iii) $\neg a \wedge \neg b \wedge c$

Viswanathan and Wallace (2003) have proposed an optimal approach to mining for Boolean functions from noisy data samples based on the minimum message length (MML) principle. The MML method has been shown to be optimal in comparison to well-known model selection methods based on guaranteed risk minimization, minimum description length (MDL) principle and cross validation after a thorough empirical evaluation with varying levels of noisy data.

## SOME RESULTS

In the first section, we have discussed the concept of conditional pattern in a database. We will use conditional patterns for synthesizing the generator of an itemset. In this section, we discuss a few results based on the discussion held in the first section.

**Lemma 1:** Let $E$ be the event that a transaction contains at least one item of the itemset $X$ in database $D$. Then the support of event $E$ in $D$ is given as follows: $S(E, D) = \sum_{Y \subseteq X,\ Y \neq \varphi} S\langle Y, X, D\rangle$.

**Proof:** The concept of support is similar to probability. We state the theorem of total probability (Sheldon, 2000) in terms of supports of relevant itemsets as follows: Let $X_1, X_2, \ldots, X_m$ be $m$ event on database $D$. Then $S\left(\bigcup_{i=1}^{m} X_i,\ D\right) = \sum_{i=1}^{m} S\left(X_i,\ D\right) -$

$$\sum_{i<j\ ;\ i,j=1}^{m} S\left(X_i \cap X_j,\ D\right) + \ldots + \left(-1\right)^{m+1} \times\ S\left(\bigcap_{i=1}^{m} X_i,\ D\right). \qquad (1)$$

From the definition of conditional pattern, we conclude that the events $\langle Y, X\rangle$ and $\langle Z, X\rangle$ are mutually exclusive, for $Y \neq Z$, and $Y, Z \subseteq X$. Thus, $S(\langle Y, X\rangle \cap \langle Z, X\rangle, D) = 0$, for $Y \neq Z$, and $Y, Z \subseteq X$. We apply the theorem of total probability to the conditional patterns with reference to itemset $X$. All the terms except the first one on the right hand side of (1) become zero, and the result follows. •

**Lemma 2:** Let $X$ and $Y$ be two itemsets in database $D$ such that $Y \subseteq X$. Let $Z = X\text{-}Y = \{a_1, a_2, \ldots, a_m\}$.

Then $S(Y, D) = \sum_{W \in \rho(Z)} S\left(\langle Y \cap W, X\rangle,\ D\right)$, where $\rho(Z) = \psi(Z) \cup \{\neg a_1 \wedge \neg a_2 \wedge \ldots \wedge \neg a_{m-1} \wedge \neg a_m\}$.

**Proof:** The proof is based on induction on $m$. Now, $S(Y, D) = S(Y \wedge a_1, D) + S(Y \wedge \neg a_1, D)$. The result is true for $m = 1$. Let the result is true for $m \leq k$. We shall show that the result is true for $m = k + 1$. For $m = k$, we have $S(Y, D) = S(Y \wedge a_1 \wedge a_2 \wedge \ldots \wedge a_k, D) + S(Y \wedge \neg a_1 \wedge a_2 \wedge \ldots \wedge a_k, D) + S(Y \wedge a_1 \wedge \neg a_2 \wedge \ldots \wedge a_k, D) + S(Y \wedge \neg a_1 \wedge \neg a_2 \wedge \ldots \wedge a_k, D) + \ldots + S(Y \wedge \neg a_1 \wedge \neg a_2 \wedge \ldots \wedge \neg a_{k-1} \wedge a_k, D) + S(Y \wedge \neg a_1 \wedge \neg a_2 \wedge \ldots \wedge \neg a_{k-1} \wedge \neg a_k, D)$, by induction hypothesis. After incorporating $a_{k+1}$, we get $S(Y, D) = S(Y \wedge a_1 \wedge a_2 \wedge \ldots \wedge a_k \wedge a_{k+1}, D) + S(Y \wedge a_1 \wedge a_2 \wedge \ldots \wedge a_k \wedge \neg a_{k+1}, D) + S(Y \wedge \neg a_1 \wedge a_2 \wedge \ldots \wedge a_k \wedge a_{k+1}, D) + S(Y \wedge \neg a_1 \wedge a_2 \wedge \ldots \wedge a_k \wedge \neg a_{k+1}, D) + \ldots + S(Y \wedge \neg a_1 \wedge \neg a_2 \wedge \ldots \wedge \neg a_{k-1} \wedge \neg a_k \wedge a_{k+1}, D) + S(Y \wedge \neg a_1 \wedge \neg a_2 \wedge \ldots \wedge \neg a_{k-1} \wedge \neg a_k \wedge \neg a_{k+1}, D)$. The result is true for $m = k + 1$.

Let us take an example. Let $Y = \{a\}$, and $X = \{a, b, c\}$. Then, $S(Y, D) = S(Y \wedge b, D) + S(Y \wedge \neg b, D) = S(Y \wedge b \wedge c, D) + S(Y \wedge b \wedge \neg c, D) + S(Y \wedge \neg b \wedge c, D) + S(Y \wedge \neg b \wedge \neg c, D)$. It validates Lemma 2.

**Lemma 3:** The framework enables in synthesizing every Boolean expression induced by a frequent itemset.

**Proof.** Let $X = \{a_1, a_2, \ldots, a_m\}$ be a frequent itemset. Then, $\psi(X) = \{a_1 \wedge a_2 \wedge \ldots \wedge a_m,\ \neg a_1 \wedge a_2 \wedge \ldots \wedge a_m,\ \ldots,\ a_1 \wedge \neg a_2 \wedge \ldots \wedge a_m,\ \neg a_1 \wedge \neg a_2 \wedge \ldots \wedge a_m,\ \ldots,\ \neg a_1 \wedge \ldots \wedge \neg a_{m-1} \wedge a_m\}$, where $|\psi(X)| = 2^m\text{-}1$. Then the members (i.e., pattern itemsets) in $\psi(X)$ form the basic building blocks for constructing an arbitrary Boolean expression induced by $X$. $\psi(X)$ induces a partition (Liu, 1985) of the set of transactions in $D$. The partition contains $2^m$ subsets of transactions in $D$.

Again, each subset of transactions corresponds to a member of $\psi(X)$, except the last one i.e., $2^m$-th subset of transactions. The last subset of transactions corresponds to the set of transactions where the Boolean expression $\neg a_1 \wedge \ldots \wedge \neg a_{m-1} \wedge \neg a_m$ is true. The support of this Boolean expression could be computed with the help of the supports of the members in $\psi(X)$. Let $E(X)$ be an arbitrary Boolean expression induced by $X$. Thus, the support of either $E(X)$ or $\neg E(X)$ could be obtained by adding supports of some members in $\psi(X)$. Thus, it is possible to synthesize the support of every Boolean expression induced by $X$. •

In particular, let $X = \{a, b, c\}$. Consider two Boolean expressions $c \vee (a \wedge \neg b)$ and $\neg a \wedge \neg c$.

The supports of the Boolean expressions in Figure 4 could be computed as follows. $S(c \vee (a \wedge \neg b), D)$ could be obtained by adding the supports of regions I, II, III, IV, and V. These regions are mutually exclusive. Each of these regions corresponds to a member of $\psi(\{a, b, c\})$. Thus, $S(E_1, D) = S(a \wedge b \wedge c, D) + S(a \wedge \neg b \wedge c, D) + S(\neg a \wedge b \wedge c, D) + S(a \wedge \neg b \wedge \neg c, D) + S(\neg a \wedge \neg b \wedge c, D)$. Also, $S(E_2, D)$ could be obtained by adding the supports of regions VI and VII. But, the region VII does not correspond to any member of $\psi(\{a, b, c\})$. Now, $S(\neg E_2, D) = S(a \wedge b \wedge c, D) + S(a \wedge \neg b \wedge c, \mathrm{D}) + S(\neg \mathrm{a} \wedge \mathrm{b} \wedge \mathrm{c}, \mathrm{D}) + S(a \wedge \neg b \wedge \neg c, D) + S(\neg a \wedge \neg b \wedge c, D) + S(a \wedge b \wedge \neg c, D)$. Therefore, $S(E_2, D) = 1 - S(\neg E_2, D)$. Thus, it validates Lemma 3.

**Lemma 4:** Let *SFIS*(*D*) be the set of frequent itemsets in database *D*. Then the number of distinct conditional patterns satisfying minimum support criterion is $\sum_{X \in \mathrm{SFIS(D)}} \left(2^{|X|} \text{ -1}\right)$.

**Proof:** The number of non-null subsets of an itemset $X$ is $2^{|X|}$ - 1, for $X \in SFIS(D)$. Each non-null subset of $X$ corresponds to a conditional pattern. The result follows.

## SYNTHESIZING GENERATORS

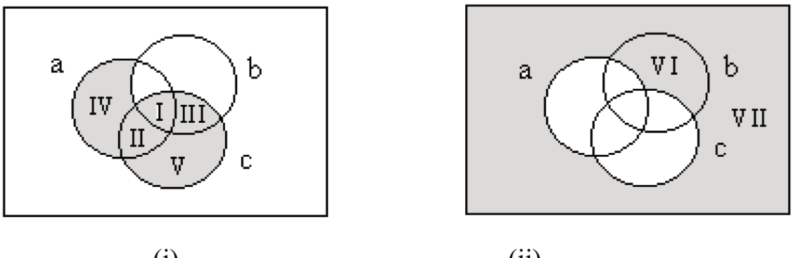
Conditional patterns are derived from the frequent itemsets in a database. Let $X$ be a frequent itemset in $D$. We shall express *csupp* $\langle Y, X, D\rangle$ in terms of the supports of the frequent itemsets in $D$, for $Y \subseteq X$. Without any loss of generality, let $X = Y \bigcup Z$, where $Z = \{a_1, a_2, \ldots, a_m\}$. The conditional support of the $i$-th conditional pattern with reference to $X$ is same as the support of the $i$-th member of $\psi(X)$, $i = 1, 2, \ldots, 2^{|X|} - 1$. Thus, the following theorem enables us to compute the supports of members of $\psi(X)$ in $D$, for all $X \in SFIS(D)$, and $|X| \geq 2$.

**Lemma 5:** Let $X$, $Y$ and $Z$ are itemsets in database $D$ such that $X = Y \bigcup Z$, where $Z = \{a_1, a_2, \ldots, a_m\}$. Then $S\langle Y, X, D\rangle = S(Y, D) -$

$$\sum_{i=1}^{m} S\left(Y \cap \left\{a_i\right\},\ D\right) +$$

$$\sum_{i < j;\ i,\ j = 1}^{m} S\left(Y \cap \left\{a_i,\ a_j\right\},\ D\right) -$$

Figure 4. *Boolean expressions $E_1(\{a, b, c\}) = c \vee (a \wedge \neg b)$ and $E_2(\{a, b, c\}) = \neg a \wedge \neg c$ represent the shaded areas of (i) and (ii) respectively*



(i)

(ii)

$$\sum_{i<j<k;\, i,j,k=1}^{m} S\left(Y \cap \left\{a_i,\ a_j,\ a_k\right\},\ D\right) + \ldots + \left(-1\right)^{m} \times S\left(Y \cap \left\{a_1,\ a_2,\ \ldots,\ a_m\right\},\ D\right) \qquad (2)$$

**Proof:** We shall prove the result using the method of induction on $m$. For $m = 1$, $X = Y \cap \left\{a_1\right\}$. Then $S\langle Y, X, D\rangle = S(Y, D) - S\left(Y \cap \left\{a_1\right\},\ D\right)$. Thus, the result is true for $m = 1$. Let the result be true for $m = p$. We shall prove that the result is true for $m = p + 1$. Let $Z = \{a_1, a_2, \ldots, a_p, a_{p+1}\}$. Due to the addition of $a_{p+1}$, the following observations are made: $S\left(Y \cap \left\{a_{p+1}\right\},\ D\right)$ is required to be subtracted, $S\left(Y \cap \left\{a_i,\ a_{p+1}\right\},\ D\right)$ is required to be added, for $1 \le i \le p$, and lastly, the term $\left(-1\right)^{p+1} \times S\left(Y \cap \left\{a_1,\ a_2,\ \ldots,\ a_{p+1}\right\},\ D\right)$ is required to be added.

Thus, $S\langle Y, X, D\rangle = S(Y, D) - \sum_{i=1}^{p+1} S\left(Y \cap \left\{a_i\right\},\ D\right) + \sum_{i<j;\, i,j=1}^{p+1} S\left(Y \cap \left\{a_i,\ a_j\right\},\ D\right) - \sum_{i<j<k;\, i,j,k=1}^{p+1} S\left(Y \cap \left\{a_i,\ a_j,\ a_k\right\},\ D\right) + \ldots + \left(-1\right)^{p+1} \times S\left(Y \cap \left\{a_1,\ a_2,\ \ldots,\ a_{p+1}\right\},\ D\right)$.

With reference to Example 1, let $Y = \{b\}$ and $X = \{a, b, c\}$. Then $S\langle Y, X, D\rangle = S(\{b\}, D) - S(\{b\} \cap \{a\}, D) - S(\{b\} \cap \{c\}, D) + S(\{b\} \cap \{a, c\}, D) = 0.8 - 0.5 - 0.3 + 0.2 = 0.2$. It tallies the value given in Table 1. Therefore, it validates Lemma 5.

## Algorithm Design

For synthesizing arbitrary Boolean queries induced by frequent itemsets in a database, we make use of an existing frequent itemset mining algorithm (Agrawal & Srikant, 1994; Han et al., 2000; Savasere al., 1995). We synthesize only the generator of Boolean expressions induced by a frequent itemset. The generator of Boolean expressions induced by the frequent itemset $X$ contains $2^{|X|}$-1 pattern itemsets. The proposed algorithm synthesizes all the members of all the generators. There are two approaches of synthesizing generators of Boolean expressions induced by frequent itemsets in a database. In the first approach, we synthesize the generator from the current frequent itemset. As soon as a frequent itemset is extracted, one could call an algorithm for synthesizing members of the corresponding generator. When a frequent itemset is found, then all the non-null subsets of this frequent itemest have already been extracted. Thus, one could synthesize all the members of the generator from the frequent itemsets extracted so far. In the second approach, one could synthesize members of the different generators after mining all the frequent itemsets. In this approach, all the frequent itemsets are processed after the mining task. These two approaches seem to be the same so far as the computational complexity is concerned. In this paper, we have followed the second approach of synthesizing members of a generator. During the process of mining frequent itemsets, the frequent itemsets of smaller size get extracted before the frequent itemsets of larger size. The frequent itemsets are kept in array *SFIS*. During the processing of current frequent itemset, all the non-null subsets are available before the current itemset in *SFIS*.

There are $2^{|X|}$ - 1 non-null subsets of itemset $X$. Each non-null subset of $X$ corresponds to a conditional pattern, and hence, it corresponds to a member of the generator of Boolean expressions induced by $X$. The subset $X$ of $X$ corresponds to a trivial conditional pattern, and it gets mined during the mining of frequent itemsets in $D$. Thus, one needs to process $2^{|X|}$ - 2 subsets of $X$.

One could view a conditional pattern as an object with the following attributes: *pattern*, *reference*, *csupp*, and *rsupp*. We use an array *CP* to store the conditional patterns in a database. The reference attribute of the $i$-th conditional pattern is accessed by the notation $CP(i)$.*reference*. Similar notations are used to access other attributes of a conditional pattern. Also, a frequent itemset could be viewed as an object with the following attributes: *itemset* and *support*. Let *N* be the number of frequent itemsets in the given database *D*. Algorithm 1 synthesizes all the members of $\psi(X)$, for $X \in SFIS(D)$.

Variable *i* keeps track of the current frequent itemset being processed. Variable *j* keeps track of number of conditional patterns generated. Using lines 3-22, each frequent itemset is processed. There are $2^{|X|}$ -1 non-null subsets of itemset *X*. Each non-null subset corresponds to a conditional pattern. The generator of *X* is synthesized using lines 6-21. Let *Y* be a subset of *X*. If $Y = X$ then the algorithm bypasses the processing of *Y* (line 8). When the algorithm synthesizes generator corresponding to a frequent itemset *X*, then it has already finished the processing of its non-null subsets. All the non-null subsets appear on or before *X* in the array *SFIS*. Thus, if a frequent itemset *X* located at position *i*, then we search for a subset of *X* from index 1 to *i* in array *SFIS*, since the array is sorted non-decreasing order on length of itemset. Thus, it justifies the condition of the while

*Algorithm 1. Synthesize generators of all frequent itemsets in the given database.*

```
procedure synthesizingGenerators (N, SFIS)
Input:
N: number of frequent itemsets in the given database
SFIS: set of frequent itemsets in the given database
Output:
Generators corresponding to the frequent itemsets
1: let i = 1;
2: let j = 0;
3: while (i ≤ N)
4:     CP(j).rsupp = SFIS(i).support;
       CP(j).reference = SFIS(i).itemset;
5:     let sum = 0;
6:     for k = 1 to (2^|SFIS(i).itemset| - 1) do
7:         tempItemset = k-th subset of SFIS(i).itemset;
8:         if (SFIS(i).itemset = tempItemset) then
9:             sum = SFIS(i).support; goto 19;
10:        end if
11:        let kk = 1;
12:        while (kk ≤ i) do
13:            if (SFIS(kk).itemset = tempItemset) then
14:                sum = sum + (-1)^(|SFIS(kk).itemset| – |tempItemset|) × SFIS(kk).support;
15:                break the while-do loop;
16:            end if
17:            kk = kk + 1;
18:        end while
19:        CP(j).csupp ← sum;
           CP(j).pattern = tempItemset;
20:        j = j + 1; i = i + 1;
21:    end for
22: end while
23: let t = 0;
24: for i = 1 to N do
25:     for k = 1 to (2^|SFIS(i).itemset| - 1) do
26:         display CP(t+k);
27:     end for
28:     t = t + 2^|SFIS(i).itemset| - 1;
29: end for
30: end procedure
```

loop at line 12. Formula (2) expresses $S\langle Y, X, D\rangle$ in terms of $S(Y\wedge Z, D)$, for all $Z \subseteq X\text{-}Y$. The coefficient of $S(Y\wedge Z, D)$ is $(-1)^{|Z|}$ in the expression of $S\langle Y, X, D\rangle$. Thus, $S(Y, Z, D) = \sum_{Z\subseteq \text{ X-Y}} \left(-1\right)^{|Z|} \times S\left(Y \wedge Z,\ D\right)$. This formula has been applied in line number 14 to calculate $S\langle Y, X, D\rangle$. We need to synthesize all the conditional patterns with reference to frequent itemsets for synthesizing generators. Lines 25-27 display the generator corresponding to *i*-the frequent itemset, $i = 1, 2, \ldots, N$. The generator corresponding to *i*-th frequent itemset contains $2^{|\text{SFIS(i).itemset}|}$ - 1 members (i.e., pattern itemsets), $i = 1, 2, \ldots, N$.

**Lemma 6:** Algorithm *synthesizingGenerators* executes in $O(N^2\times 2^p)$ time, where $N$ and $p$ are the number of frequent itemsets and average size of frequent itemsets in the database, respectively.

**Proof:** The *while-loop* at line 3 repeats $N$ times. The *for-loop* at line 6 repeats $2^p$-1 times. Also, the *while-loop* at line 12 repeats maximum $N$ times. Thus, the time complexity of lines 3-22 is $O(N^2\times 2^p)$. The time complexity of lines 24-29 is $O(N\times 2^p)$. Therefore, the time complexity of algorithm *synthesizingGenerators* is $O(N^2\times 2^p)$. •

## Synthesizing First k Boolean Queries Induced by Top p Frequent Itemsets

In this section, we discuss the Boolean expressions that we mine in the fifth section. A Boolean expression could be synthesized by the members of the corresponding generator. We classify the frequent itemsets in a database into different categories. The frequent itemsets of the same size are put in the same category. We sort the frequent itemsets of each category in non-increasing order by support and top frequent itemsets in each category are considered for synthesis. We perform experiments for synthesizing first *k* Boolean expressions induced by top *p* frequent itemsets of each category.

**Example 2:** Let $\{a, b\}$ and $\{a, b, c\}$ be two frequent itemsets in $D$ of size 2 and 3, respectively. We would like to determine first *k* Boolean expressions induced by $\{a, b\}$ and $\{a, b, \text{c}\}$. Let $E_{\text{ij}}$ be the *j*-th Boolean expression induced by the frequent itemset of size $i$, $j = 1, 2, \ldots, 2^i - 1$, and $i = 2, 3$. The truth tables for the first six Boolean expressions are given Table 3.

First six Boolean expressions are based on the items of $\{a, b\}$, and the rest of the Boolean expressions are based on the items of $\{a, b, c\}$. The algebraic expressions of first six Boolean expressions are given as follows: $E_{21}(a, b) = 0$,

*Table 3. Truth tables for the first six Boolean expressions induced by {a, b} and {a, b, c}*

| *a* | *b* | *c* | $E_{21}$ | $E_{22}$ | $E_{23}$ | $E_{24}$ | $E_{25}$ | $E_{26}$ | $E_{31}$ | $E_{32}$ | $E_{33}$ | $E_{34}$ | $E_{35}$ | $E_{36}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |

$E_{22}(a, b) = a \wedge b$, $E_{23}(a, b) = a \wedge \neg b$, $E_{24}(a, b) = a$, $E_{25}(a, b) = \neg a \wedge b$, $E_{26}(a, b) = b$; $E_{31}(a, b, c) = 0$, $E_{32}(a, b, c) = a \wedge b \wedge c$, $E_{33}(a, b, c) = a \wedge b \wedge \neg c$, $E_{34}(a, b, c) = a \wedge b$, $E_{35}(a, b, c) = a \wedge \neg b \wedge c$, $E_{36}(a, b, c) = a \wedge c$. We express $E_{ij}$s in terms of members of the generator. Boolean expressions $E_{22}$, $E_{23}$, and $E_{25}$ have already been expressed in terms of the members of the generator. We need not compute $E_{21}$ and $E_{31}$, since $E_{21}(a, b) = E_{31}(a, b, c) = 0$. $E_{24}(a, b) = (a \wedge b) \vee (a \wedge \neg b)$, and $E_{26}(a, b) = (a \wedge b) \vee (\neg a \wedge b)$. Also, $E_{34}(a, b, c) = (a \wedge b \wedge c) \vee (a \wedge b \wedge \neg c)$, and $E_{36}(a, b, c) = (a \wedge b \wedge c) \vee (a \wedge \neg b \wedge c)$. Expressions $E_{32}$, $E_{33}$ and $E_{35}$ have already been expressed in terms of the members of the generator.

## EXPERIMENTS

We have carried out several experiments to study the effectiveness of the framework. All the experiments have been implemented on a 1.6 GHz Pentium IV with 256 MB of memory using visual C++ (version 6.0) software. We present the experimental results using three real and one synthetic datasets. The dataset *retail* (Frequent itemset mining dataset repository) is obtained from an anonymous Belgian retail supermarket store. The datasets *BMS-Web-Wiew-1* and *BMS-Web-Wiew-2* can be found from KDD CUP 2000 (Frequent itemset mining dataset repository). The dataset *T10I4D100K* (Frequent itemset mining dataset repository) was generated using the generator from IBM Almaden Quest research group. We present some characteristics of these datasets in Table 4.

Let *NT*, *AFI*, *ALT*, and *NI* denote the number of transactions, average frequency of an item, average length of a transaction and the number of items in the corresponding dataset, respectively. For the purpose of synthesizing Boolean expressions, we have implemented apriori algorithm (Agrawal & Srikant, 1994), since it is simple and easy to implement. In Table 5, Table 6, Table 7, Table 8, Table 9, and Table 10, we present first six Boolean expressions induced top five frequent itemsets from *retail*, *BMS-Web-Wiew-1*, *BMS-Web-Wiew-2* and *T10I4D100K*, respectively.

The Boolean functions $E_{22}$ and $E_{32}$ are not shown, since they are all equal to 0. The Boolean expressions induced by frequent itemsets of size one are not studied here. *BMS-Web-Wiew-1* and *BMS-Web-Wiew-2* do not report any frequent itemsets of size greater than two. *T10I4D100K* reports only one frequent itemset of size 3. We observe that the proposed framework is simple and elegant. It enables us to synthesize arbitrary Boolean queries induced by frequent itemsets in a dataset.

Also, we have conducted experiments to study the relationship between the size of a dataset and the execution time required for mining generators. The execution time increases as the number of transactions contained in a dataset increases. We observe this phenomenon in Figures 5 and 6.

We have also conducted experiments to find the execution time for synthesizing generators in a dataset. The time required (only) for synthesizing generators for each of the above datasets is 0 millisecond at the respective values of $\alpha$ as shown in Tables 5, 6, 7, and 8.

Also, we have conducted experiments to study the relationship between the size of a dataset and the number of generators of Boolean expressions induced by frequent itemsets of

*Table 4. Dataset characteristics*

| Dataset | *NT* | *ALT* | *AFI* | *NI* |
|---|---|---|---|---|
| *retail* | 88,162 | 11.3058 | 99.6738 | 10,000 |
| *BMS-Web-Wiew-1* | 1,49,639 | 2.0000 | 155.7118 | 1,922 |
| *BMS-Web-Wiew-2* | 3,58,278 | 2.0000 | 7,165.5600 | 100 |
| *T10I4D100K* | 1,00,000 | 11.1023 | 1,276.1241 | 870 |

*Table 5. First six Boolean expressions induced by top five frequent itemsets of size 2 in retail at α = 0.05*

| frequent itemset | $S(E_{22}, D)$ | $S(E_{23}, D)$ | $S(E_{24}, D)$ | $S(E_{25}, \mathbf{D})$ | $S(E_{26}, D)$ |
|---|---|---|---|---|---|
| {39, 48} | 0.3306 | 0.2562 | 0.5868 | 0.1582 | 0.4888 |
| {39, 41} | 0.1295 | 0.4573 | 0.5868 | 0.0422 | 0.1717 |
| {38, 39} | 0.1173 | 0.0603 | 0.1776 | 0.4694 | 0.5868 |
| {41, 48} | 0.1023 | 0.0694 | 0.1717 | 0.3865 | 0.4888 |
| {32, 39} | 0.0959 | 0.0793 | 0.1752 | 0.4909 | 0.5868 |

*Table 6. First six Boolean expressions induced by top five frequent itemsets of size 3 in retail at α = 0.05*

| frequent itemset | $S(E_{32}, D)$ | $S(E_{33}, D)$ | $S(E_{34}, D)$ | $S(E_{35}, D)$ | $S(E_{36}, D)$ |
|---|---|---|---|---|---|
| {39,41,48} | 0.0836 | 0.0459 | 0.1295 | 0.2470 | 0.3306 |
| {38,39,48} | 0.0692 | 0.0481 | 0.1173 | 0.0209 | 0.0901 |
| {32,39,48} | 0.0613 | 0.0346 | 0.0959 | 0.0299 | 0.0911 |
| {1,39,48} | 0.0449 | 0.0215 | 0.0663 | 0.0170 | 0.0618 |
| {5,39,48} | 0.0432 | 0.0197 | 0.0629 | 0.0162 | 0.0594 |

*Table 7. First six Boolean expressions induced by top five frequent itemsets of size 2 in BMS-Web-Wiew-1 at α = 0.01*

| frequent itemset | $S(E_{22}, D)$ | $S(E_{23}, D)$ | $S(E_{24}, D)$ | $S(E_{25}, D)$ | $S(E_{26}, D)$ |
|---|---|---|---|---|---|
| {5, 7} | 0.0139 | 0.2353 | 0.2491 | 0.2012 | 0.2151 |
| {1, 5} | 0.0137 | 0.1761 | 0.1899 | 0.2355 | 0.2491 |
| {3, 5} | 0.0131 | 0.1953 | 0.2083 | 0.2361 | 0.2491 |
| {5, 9} | 0.0129 | 0.2363 | 0.2491 | 0.2014 | 0.2142 |
| {1, 7} | 0.0124 | 0.1774 | 0.1899 | 0.2027 | 0.2151 |

*Table 8. First six Boolean expressions induced by top five frequent itemsets of size 2 in BMS-Web-Wiew-2 at α = 0.009*

| frequent itemset | $S(E_{22}, D)$ | $S(E_{23}, D)$ | $S(E_{24}, D)$ | $S(\mathbf{E}_{25}, \mathbf{D})$ | $S(E_{26}, D)$ |
|---|---|---|---|---|---|
| {1, 3} | 0.0236 | 0.1695 | 0.1932 | 0.1710 | 0.1946 |
| {1, 7} | 0.0229 | 0.1702 | 0.1932 | 0.1741 | 0.1970 |
| {3, 7} | 0.0228 | 0.1719 | 0.1946 | 0.1743 | 0.1970 |
| {3, 5} | 0.0220 | 0.1726 | 0.1946 | 0.1572 | 0.1793 |
| {1, 9} | 0.0220 | 0.1712 | 0.1932 | 0.1512 | 0.1732 |

*Table 9. First six Boolean expressions induced by top five frequent itemsets of size 2 in T10I4D100K at α = 0.01*

| frequent itemset | $S(E_{22}, D)$ | $S(E_{23}, D)$ | $S(E_{24}, D)$ | $S(E_{25}, D)$ | $S(E_{26}, D)$ |
|---|---|---|---|---|---|
| {217, 346} | 0.0134 | 0.0405 | 0.0539 | 0.0214 | 0.0347 |
| {789, 829} | 0.0119 | 0.0335 | 0.0454 | 0.0690 | 0.0809 |
| {368, 829} | 0.0119 | 0.0665 | 0.0785 | 0.0690 | 0.0809 |
| {368, 682} | 0.0119 | 0.0665 | 0.0785 | 0.0319 | 0.0438 |
| {39, 825} | 0.0119 | 0.0307 | 0.0426 | 0.0237 | 0.0356 |

*Table 10. First six Boolean expressions induced by frequent itemsets of size 3 in T10I4D100K at α = 0.01*

| frequent itemset | $S(E_{32}, D)$ | $S(E_{33}, D)$ | $\mathbf{S(E_{34}, D)}$ | $\mathbf{S(E_{35}, D)}$ | $\mathbf{S(E_{36}, D)}$ |
|---|---|---|---|---|---|
| {39, 704, 825} | 0.0104 | 0.0004 | 0.0111 | 0.0015 | 0.0119 |

*Figure 5. Execution time versus the number of transactions from retail at α = 0.05*

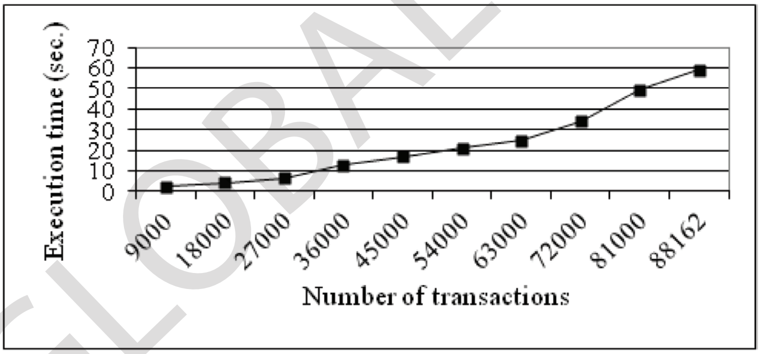
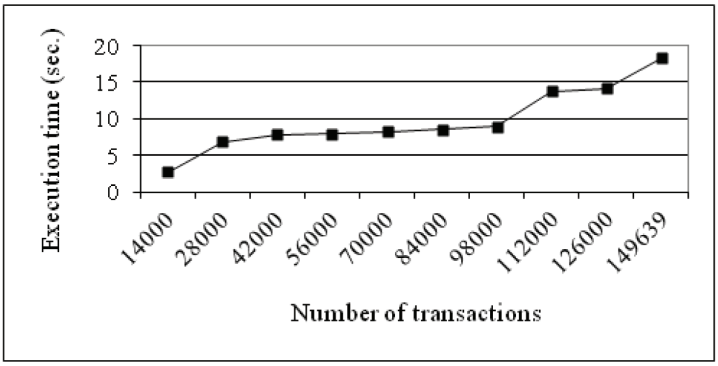


*Figure 6. Execution time versus the number of transactions from BMS-Web-Wiew-1 at α = 0.01*



size greater than or equal to 2. The experiments are conducted on datasets *retail* and *BMS-Web-Wiew-1*. The results of the experiments are shown in Figures 7 and 8. From Figures 7 and 8, we could conclude that there is no universal relationship between the size of the dataset and the number of generators.

We have also conducted experiments for finding the number of generators corresponding to frequent itemsets of size greater than or equal to 2 in a dataset at a given $\alpha$. The number of generators in a dataset decreases as α increases. We observe this phenomenon in Figures 9 and 10.

## AN APPLICAION: DATA MINING APPROACH TO REDUCE QUERY PROCESSING TIME

Query processing is a significant task performed by a database system. To answer a query from a large database we need to access data from the secondary memory. Thus, the processing time is dependent on accessing the secondary memory. Query processing in a database has been well studied (Freytag et al., 1993; Kim et al., 1985). The goal of this section is to propose a model for processing queries using patterns in a database.

Let $I(D)$ and $FI(D)$ be the set of items and set of frequent items in database $D$, respectively. Let $f(x_1, x_2, \ldots, x_p)$ be an arbitrary Boolean query. Most of the cases users might be interested in mining a Boolean function of frequent items. In Lemma 3, we have shown that every Boolean function induced by frequent items could be synthesized by relevant frequent itemsets. In this case, one could keep all the frequent itemsets in the main memory. Thus, such queries could be answered quickly, since it is not required to access secondary memory. Such queries could be called *frequent queries*. Other queries could be answered by accessing data from the secondary storage. They could be called *infrequent queries*. No work has been reported on answering queries after the mining process. In this paper, we propose a model of answering an arbitrary Boolean query during the post processing of patters.

At the initial stage the database is mined, and all the frequent itemsets are stored in the main memory for synthesizing frequent queries. A query is processed for determining whether it is frequent. A query $f(x_1, x_2, \ldots, x_p)$ is called frequent if $x_i \in FI(D)$, for all $i = 1, 2, \ldots, p$.

There are three stages of processing a frequent query. At the first stage, a query is pursed to determine whether it is frequent. At the second stage, a query is expressed using itemsets. Let $IS(f)$ be the set of itemsets that are required to synthesize query $f$. If $f$ is frequent then each member of $IS(f)$ is searched in the main memory and finally, the support of $f$ is synthesized using supports of itemsets in $IS(f)$. Thus, at the third stage, a query $f$ is synthesized using itemsets in $IS(f)$. But in case of infrequent

*Figure 7. Number of generators versus the number of transactions from retail at α = 0.05*

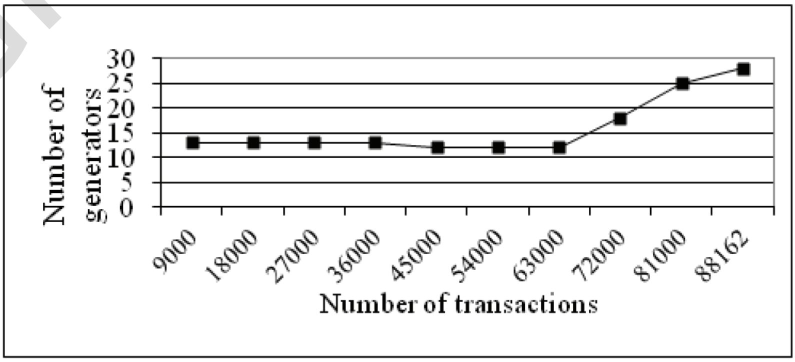


*Figure 8. Number of generators versus the number of transactions from BMS-Web-Wiew-1 at α = 0.01*

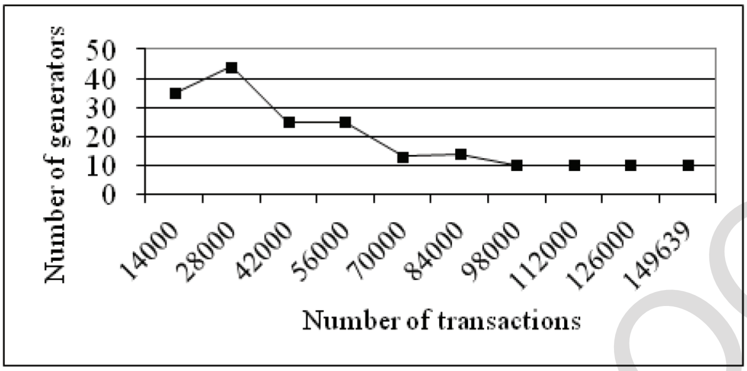


*Figure 9. Number of generators versus α for retail*

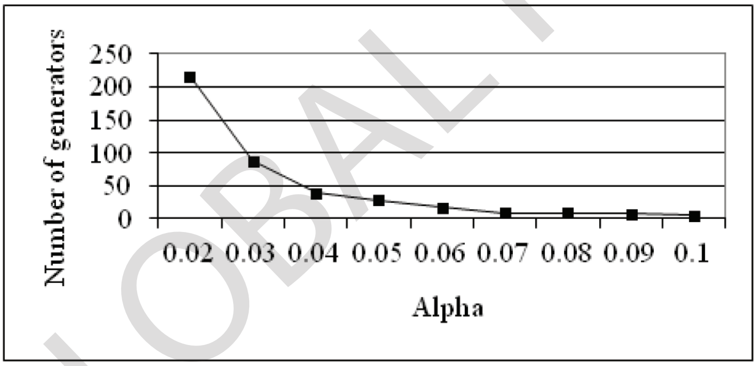
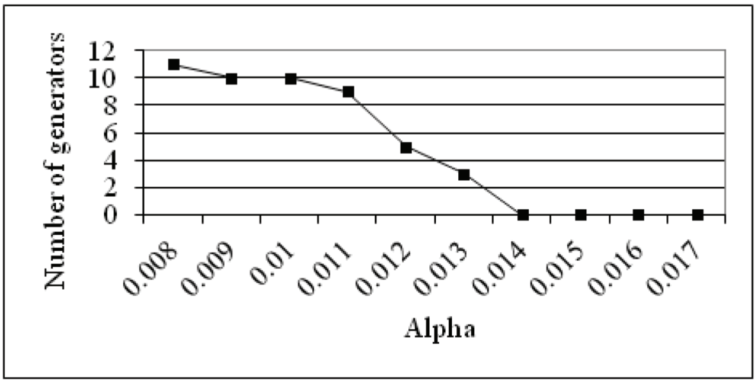


*Figure 10. Number of generators versus α for BMS-Web-Wiew-1*



queries, there are four stages in the synthesizing process. Let *IIS*(*f*) be the set of infrequent itemsets that are also required to synthesize infrequent query *f*. Then for infrequent query *f*, the itemsets in *IS*(*f*) – *IIS*(*f*) have already been mined. Thus, at the third stage, the database *D* is mined for itemsets in *IIS*(*f*). One could call it as a constrained data mining, where we are interested in mining a specific set of itemsets. Thus, at the fourth stage, infrequent query *f* is synthesized using itemsets in sets *IS*(*f*) – *IIS*(*f*) and *IIS*(*f*).

There is a significant saving of processing time in finding answer of a query, since a large number of queries are frequent. Moreover, a given query may consist of a significant number of frequent items. Thus, it reduces the secondary memory access. The proposed technique could not be an alternative to the conventional query optimization techniques. Rather, it could be a query preprocessing technique in a knowledge and information system. The conventional technique could be applied as well after the proposed query preprocessing. Overall, the average query processing time could be significantly reduced using the proposed approach.

The rest of the section is organized as follows. In the first subsection of the sixth section, we discuss a method of synthesizing support of a Boolean query. In the second subsection of that section, we discuss a constrained data mining. A model of reducing query processing using data mining approach is discussed in the third subsection.

## Synthesizing Support of a Boolean Query

Let *E* be a Boolean expression on items in database *D*. Support of *E* in *D* could be defined as the fraction of transactions in *D* such that the Boolean expression *E* is true for each of these transactions. In a database management system, a query could be placed interactively by a user. Initially, the query is parsed and checked for its syntax. Then the support of the Boolean query could be expressed by sum of the supports of relevant frequent itemsets as illustrated by Example 3.

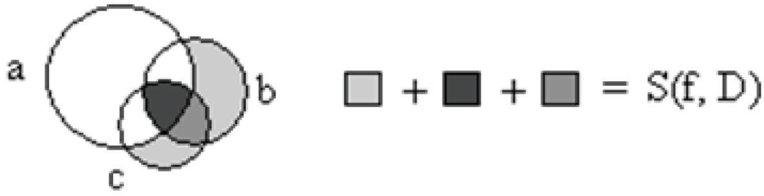
**Example 3:** Consider the Boolean query $f(a, b, c) = (\neg a \wedge (b \vee c)) \vee (b \wedge c)$. The support of shaded region of Figure 11 is required to be determined.

There are two cases viz., (i) All the items *a*, *b*, and *c* are frequent, and (ii) At least one of items among *a*, *b*, and *c* is not frequent. Before discussing the above cases, first we need to express support of given Boolean function in terms of supports of relevant itemsets. Using Figure 11, support of *f* could be computed as follows:

$S(f, D) = S(\{b\}, D) + S(\{c\}, D) - S(\{a, b\}, D) - S(\{a, c\}, D) - S(\{b, c\}, D) + 2 \times S(\{a, b, c\}, D)$ (3)

In case of (i), *S*(*f*, *D*) could be obtained by the available supports in the main memory. In case of (ii), let the item *b* be infrequent. Then, the supports of itemsets {*b*}, {*a*, *b*}, {*b*, *c*} and

*Figure 11. Sum of shaded regions corresponds to the support of f(a, b, c)*



{*a*, *b*, *c*} are not available in the main memory. Thus *IIS*(*f*) = { {*b*}, {*a*, *b*}, {*b*, *c*}, {*a*, *b*, *c*} }. Supports of itemsets in *IIS*(*f*) are required to be mined. We mine database *D* using input *IIS*(*f*). This type of mining could be called as a constrained data mining. •

Let S-CVRT(*f*) be an algorithm that takes a Boolean function *f* as input and calculates the support of *f* in terms of supports of relevant itemsets in the database.

It could be possible to write an algorithm S-CVRT (see Figure 12). We have discussed this issue in Lemma 3. Using probabilistic results (Papoulis, 1984), it could be possible write an algorithm for such converter. In the sixth section's second subsection, we will discuss how apriori algorithm could be used to mine only a set of specific (infrequent) itemsets.

## Mining a Set of Specific Itemsets in a Database

We present an apriori-based algorithm, *SSIMining*, for mining a set of specific itemsets (SSI) in the given database. In this algorithm, we use a two-dimentional array *IIS*. The first row contains a set of infrequent itemsets. After the end of mining, the second row would contain the corresponding supports. The algorithm (Algorithm 2) is presented below.

Initially, all the 1-itemsets are mined from *D*. Then all the 2-itemsets are mined using 1-itemsets in *IIS*, and so on. Symbol |*IIS*(1, *j*)| stands for the number of items in itemset *IIS*(1, *j*), for a *j* = 1, 2, … . Constrained data mining algorithm *SSIMining* is faster, since we mine itemsets at the current level based on itemsets at the previous level and using only those that are present in *IIS*.

## A Model of Query Processing

When a query is submitted into a database management system, the query is passed through an algorithm, called PARSER. It checks whether the query is syntactically valid. Algorithm for PARSER (see Figure 13) is beyond the scope of this paper.

Using theory of compilation (Aho et al., 2003), it is possible to write an algorithm for such parser. A valid query is passed to the next phase. In this phase, we use algorithm S-CVRT for representing it using itemsets in the databases. If *IIS*(*f*) = $\phi$ then the query could be answered using the supports of frequent itemsets in the database. Before querying the database, we mine all the frequent itemsets. The frequent itemsets are kept in the main memory. If *IIS*(*f*) ≠ $\phi$ then the we apply constrained mining algorithm C-MINE to extract necessary itemsets from the database. Finally, the query is synthesized using the supports of relevant itemsets in the database. Based on the above discussion, we present here a model of synthesizing query in a transactional database.

In Figure 14, each rectangle represents an algorithm. Initially, we mine database *D*, and obtain the supports of itemsets in *IS*(*f*)-*IIS*(*f*). If we get at least one infrequent itemset from query *f* then we need to mine the database to determine supports of itemsets in *IIS*(*f*). Finally, we synthesize the support of *f* using supports of itemsets in *IIS*(*f*) and *IS*(*f*)-*IIS*(*f*).

## CONCLUSION

Frequent itemsets determine the major characteristics of a database. One could analyze the characteristics of a database in more detail by mining arbitrary Boolean expressions. This paper proposes a framework for synthesizing generators of Boolean expressions induced by frequent itemsets. The generators enable us in synthesizing arbitrary Boolean expressions induced by the frequent itemsets. The framework also enables in synthesizing Boolean expressions containing NOT operator. It is a simple and elegant technique. There is no need to introduce a specific framework for a specific type of Boolean expressions. The proposed framework is effective and promising. We have presented an application of the proposed framework. In this application, we have shown how one could reduce average query processing time in a database. In Figure 14, we have proposed

*Figure 12. S-CVRT algorithm*

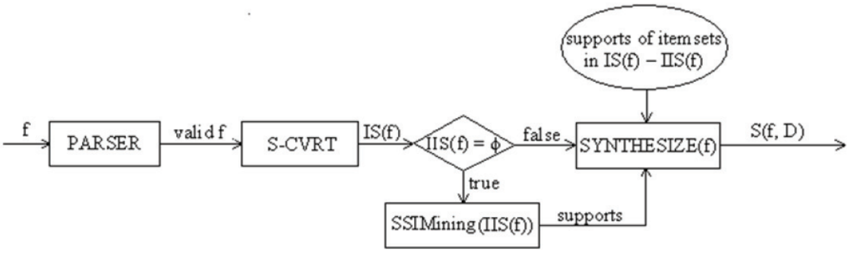
*Algorithm 2. Mine a set of specific itemsets in a given database.*

```
procedure SSIMining (IIS)
Input:
First row of array IIS contains set of itemsets
Output:
Second row of array IIS contains supports of infrequent itemsets
sort itemsets in IIS based on size of an itemset;
let the number of itemset be m;
let level = 1; let index = 1;
while (level ≤ |IIS(1, m)|) do
    mine database D using itemsets at the previous level;
    store all itemsets at the current level into array temp;
    while (level = |IIS(1, index)|) do
        search IIS(1, index) into temp;
        update IIS(2, index) by the value found in temp;
        increase index by 1;
    end while
    increase level by 1;
end while
end procedure
```

*Figure 13. Checking validity of a query*

f
PARSER
valid f

*Figure 14. A model of synthesizing query f in database D*



a model of synthesizing query in a database. In future, we shall implement algorithms S-CVRT, PARSER, and SYNTHESIZE to test this model on real and synthetic databases.

## ACKNOWLEDGMENT

The author acknowledges the Editor-in-Chief of the journal, Professor John Wang, and the anonymous reviewers for their indispensable input that improved the paper significantly.

## REFERENCES

Adhikari, A., & Rao, P. R. (2008). Efficient clustering of databases induced by local patterns. *Decision Support Systems*, *44*(4), 925–943. doi:10.1016/j.dss.2007.11.001.

Adhikari, A., & Rao, P. R. (2008). Mining conditional patterns in a database. *Pattern Recognition Letters*, *29*(10), 1515–1523. doi:10.1016/j.patrec.2008.03.005.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD Conf. Management of Data* (pp. 207-216).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the International Conference on Very Large Data Bases* (pp. 487-499).

Aho, A., Sethi, R., & Ullman, J. D. (2003). *Compilers: Principles, techniques, and tools*. Pearson Education.

Antonie, M.-L., & Zaïane, O. R. (2004). Mining positive and negative association rules: An approach for confined rules. In *Proceedings of PKDD* (pp. 27-38).

Baralis, E., & Psaila, G. (1999). Incremental refinement of mining queries. In *Proceedings of the 1$^{st}$ DaWaK Conference* (pp. 173-182).

Bonchi, F., & Lucchese, C. (2005). Pushing tougher constraints in frequent pattern mining. In *Proceedings of PAKDD* (pp. 114-124).

Cheung, D. W., Han, J., Ng, V., & Wong, C. Y. (1996). Maintenance of discovered association rules in large databases: An incremental updating technique. In *Proceedings of the 12th ICDE* (pp. 106-114).

*FIMI.* (2004). *Frequent itemset mining dataset repository.* Retrieved from http://fimi.cs.helsinki.fi/data.

Freytag, J., Maier, D., & Vossen, G. (1993). *Query processing for advanced database systems*. Morgan Kauffman.

Han, J., Pei, J., & Yiwen, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings ACM SIGMOD Conference Management of Data* (pp. 1-12).

Kim, W., Reiner, D. S., & Batory, D. S. (1985). *Query processing in database systems*. Springer. doi:10.1007/978-3-642-82375-6.

Liu, C. L. (1985). *Elements of discrete mathematics* (2nd ed.). McGraw-Hill.

Mediavilla, F. M. A., & Lee, H.-M. (2009). Approximate queries on distributed data marts. *International Journal of Information and Decision Sciences*, *1*(4), 366–381. doi:10.1504/IJIDS.2009.027757.

Meo, R. (2003). Optimization of a language for data mining. In *Proceedings of the ACM Symposium on Applied Computing - Data Mining Track* (pp. 437-444).

Morzy, T., Wojciechowski, M., & Zakrzewicz, M. (2000). Materialized data mining views. In *Proceedings of the 4th PKDD Conference* (pp. 65-74).

Muhonen, J., & Toivonen, H. (2006). Closed non-derivable itemsets. In *Proceedings of PKDD* (pp. 601-608).

Papoulis, A. (1984). *Probability, random variables, and stochastic processes* (2nd ed.). McGraw-Hill.

Pavlov, D., Mannila, H., & Smyth, P. (2000). Probabilistics models for query approximation with large sparse binary data sets. In *Proceedings of Sixteenth Conference on Uncertainty in Artificial Intelligence* (pp. 465-472).

Pei, J., & Han, J. (2000). Can we push more constraints into frequent pattern mining? In *Proceedings of KDD* (pp. 350-354).

Savasere, A., Omiecinski, E., & Navathe, S. (1995). An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21$^{st}$ International Conference on Very Large Data Bases* (pp. 432-443).

Sheldon, R. M. (2000). *Introduction to probability models* (7th ed.). Academic Press.

Shima, Y., Mitsuishi, S., Hirata, K., Harao, M., Suzuki, E., & Arikawa, S. (2004). Extracting minimal and closed monotone dnf formulas. In *Proceedings of International Conference on Discovery Science, 3245*, 298-305.

Sun, L., & Li, Y. (2009). Using usage control to access XML databases. *International Journal of Information Systems in the Service Sector*, *1*(3), 32–44. doi:10.4018/jisss.2009070102.

Viswanathan, M., & Wallace, C. (2003). An optimal approach to mining Boolean functions from noisy data. In *Proceedings of IDEAL* (pp. 717-724).

Wu, X., Wu, Y., Wang, Y., & Li, Y. (2005). Privacy-aware market basket data set generation: A feasible approach for inverse frequent set mining. In *Proceedings of SIAM International Conference on Data Mining*, (pp. 103-114).

Wu, X., Zhang, C., & Zhang, S. (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems*, *22*(3), 381–405. doi:10.1145/1010614.1010616.

Wu, X., Zhang, C., & Zhang, S. (2005). Database classification for multi-database mining. *Information Systems*, *30*(1), 71–88. doi:10.1016/j.is.2003.10.001.

Yu, C. T., & Meng, W. (1997). *Principle of database query processing for advanced applications*. Morgan Kaufmann.

Zhao, L., Zaki, M. J., & Ramakrishnan, M. (2006). BLOSOM: A framework for mining arbitrary Boolean expressions. In *Proceedings of KDD*, (pp. 827-832).

*Animesh Adhikari is an Associate Professor of Computer Science at the Department of Computer Science, S. P. Chowgule College, Margao, Goa, India. He obtained his MTech and PhD degrees, both in Computer Science, from Indian Statistical Institute and Goa University, respectively. His areas of interest include data mining and knowledge discovery, decision support systems, database systems, and artificial intelligence. He has published one research monograph, nine international journal papers and five international conference papers.*

# A Knowledge Mining Approach for Effective Customer Relationship Management

*Fatudimu Ibukun Tolulope, Department of Computer and Information Sciences, Covenant University, Ota, Nigeria*

*Charles Uwadia, Department of Computer Science, University of Lagos, Lagos, Nigeria*

*C. K. Ayo, Department of Computer and Information Sciences, Covenant University, Ota, Nigeria*

## ABSTRACT

*The problem of existing customer relationship management (CRM) system is not lack of information but the ability to differentiate useful information from chatter or even disinformation and also maximize the richness of these heterogeneous information sources. This paper describes an improved text mining approach for automatically extracting association rules from collections of textual documents. It discovers association rules from keyword features extracted from the documents. The main contributions of the technique are that, in selecting the most discriminative keywords for use in association rules generation, the system combines syntactic and semantic relevance into its Information Retrieval Scheme which is integrated with XML technology. Experiments carried out revealed that the extracted association rules contain important features which form a worthy platform for making effective decisions as regards customer relationship management. The performance of the improved text mining approach is compared with existing system that uses the GARW algorithm to reveal a significant reduction in the large itemsets, leading to reduction in rules generated to more interesting ones due to the semantic analysis component being introduced. Also, it has brought about reduction of the execution time, compared to the GARW algorithm.*

*Keywords:* *Association Rule Mining, Chatter, Customer Relationship Management (CRM), Evaluation, Generating Association Rules Based on Weighting Scheme (GARW) Algorithm, Text Mining*

## INTRODUCTION

Presently, information processing is gradually moving towards semi structured or unstructured data management (Kernochan, 2006). Most data-mining research assumes that the information to be "mined" is already in the form of a relational database. Unfortunately, according to (Raymond et al., 2006), in many applications, available electronic information is in the form of unstructured data. Text mining is a technology that makes it possible to

DOI: 10.4018/ijkbo.2013040105

discover patterns and trends semi automatically from huge collections of unstructured text. It is based on technologies such as natural language processing, information retrieval, information extraction, and data mining (Andrea et al., 2010). The term text mining was coined to describe tools used to manage textual information. Text mining, known as knowledge discovery in textual databases (Ah-Hwee, 1999) can also be defined as the application of data mining techniques to automated discovery of useful or interesting knowledge from unstructured text (Han et al., 2000). It allows the creation of a technology that combines a human's linguistic capabilities with the speed and accuracy of a computer. Text mining aims at employing technology to analyze more detailed information in the content of each document and to extract interesting information that can be provided only by multiple documents viewed as whole, such as trends and significant features that may be a trigger to useful actions and decision making (Nasukawa et al., 2001).

Text data mining is a much more complex task than data mining (Ah-Hwee, 1999), because it involves text data that is inherently unstructured and fuzzy. Knowledge discovery in text can be broadly classified into two main phases. Firstly, transformation of (free-form) text documents into an internal or intermediate form and secondly Text mining, which is called knowledge distillation and it is the phase that deduces patterns or knowledge from the intermediate form.

This paper aims at studying particular features of texts, identifying patterns that may be used for making relevant business decision and discussing the tools that may be used for such purpose. In discussing the tools, text mining technique that is based on modification of the GARW algorithm is described (Hany, Dietmar, Nabil, & Fawzy, 2007). Text documents were selected from questionnaires which were administered in order to elicit information towards effective customer relationship management in the mobile phone manufacturing industry.

The Motivations for choosing this domain are that:

- Reports on one study showing customer service channels used by 60 firms revealed that information is stored most times in unstructured form (Strauss, El-Ansary, & Frost, 2006).
- There exists a challenge within the field of customer relationship management and competitive Intelligence which is not lack of information but the ability to differentiate useful information from chatter or even disinformation and also maximize the richness of these heterogeneous information sources (Solomon et al., 2003).

In Customer relationship management, information is the raw material for decision making (Graham, John, & Nigel, 2004). Effective market decisions are therefore based on sound information and the decisions are not better than the information on which they are based. Information is therefore the lubricant of Business Intelligence. The more information a firm has, the better the value it can provide to each customer and the better the prospects in terms of more accurate, timely and relevant offerings (Strauss et al., 2006).

In this paper therefore we present the use of association rule in Text Mining. These association rules highlight correlations between keywords in the text. Association rules is appropriate for the area of application because they are easy to understand and interpret for top management staff who might be the user of such a system.

The rest of the paper is organized as follows. The second section presents a review of related work, the third section presents the text mining system architecture, implementation and discussion are presented in the fourth section. The fifth section describes the evaluation results while the sixth and seventh sections are the conclusion and the future work respectively.

## RELATED WORK

According to (Feldman et al., 1995; Feldman et al., 1996), association rules have been mined from manually assigned keywords. This method has the following disadvantages; its time consuming, subjected to discrepancies and the textual sources are constrained to those that have the keywords predetermined. (Rajman et al., 1997) presented two examples of text mining tasks, association extraction and prototypical document extraction, along with several related NLP techniques. In the case of association extraction task, association rules were extracted from a collection of indexed documents. In their work on mining association rule form biomedical text, (Berardi, Malerba, Marinelli, Leo, Loglisci, & Scioscia, 2005) performed entity extraction using BioTeKS which aims at both identifying the location of an entity in a text and categorize it according to the standard MeSH (Medical Subject Headings) taxonomy. This makes the application restricted to only the medical domain. In their work on the survey of basic concepts in the area of text data mining and some of the methods used in order to elicit useful knowledge from collections of textual data, the authors suggested that there has been some minor attempts to use (partially or fully) structured textual documents such as HTML or XML documents in order develop text mining systems (Jan et al., 1999).

There are also some approaches to text mining with information extraction as the text preprocessing phase, which inherits the generational problems of information extraction systems which include uncertainty of extracted features. In addition, Information Retrieval (IR) techniques have widely used the "bag-of-words" model (Baeza-Yates et al., 1999; Raymond et al., 2005) for tasks such as document matching, ranking, and clustering. Some of the approaches include that of (Shenzhi et al., 2004) where an algorithm that learns rules and extracts entities from unstructured textual data was developed. In Michal, Martin, Emil, Zoltan, and Ladislav (2004), semi-automatic ontology based text annotation (OnTeA) tool was used to analyze document or text using regular expression patterns and detects equivalent semantics elements according to the defined domain ontology which can then be used as input to a text mining system, again, this approach is domain specific.

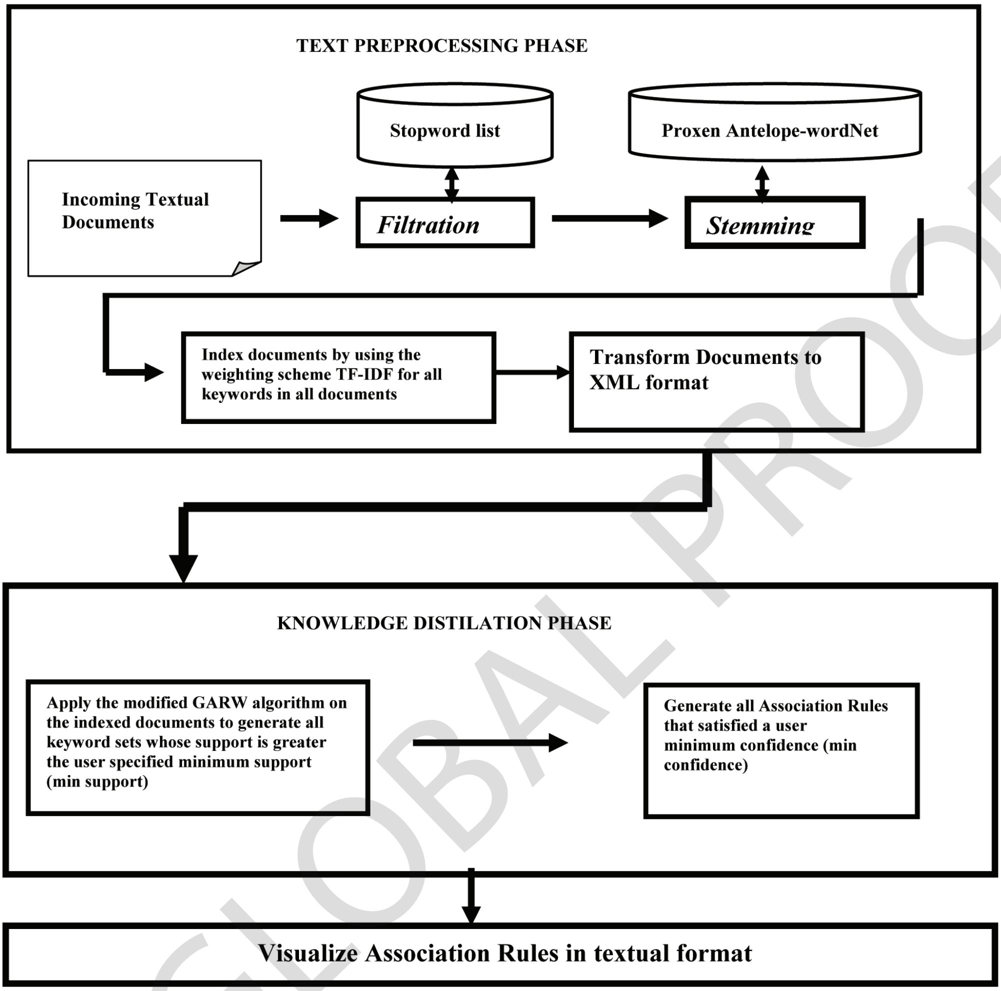
In this work, the focus is on extraction of association rules based on extracted keywords that are generated as a result of a text preprocessing phase that references a machine readable dictionary (wordNet) which makes such a system suitable for any domain of application. The process of mining association rules in temporal document collection and performing the various steps in the temporal text mining process are described in Norvag, Eriksen, and Skgstad (2006). In Chen et al. (2001) the authors presented a text mining technique that discovers association rules from documents for a particular user. It derives a user's background knowledge from his background documents, and exploits such knowledge in the form of association rules they also used TF-IDF (Term Frequency, Inverse Document Frequency) to select significant noun phrases from each target document. In Liu, Navathe, Pivoshenko, Dasigi, Dingledine, and Ciliax (2006) the authors evaluate the effectiveness of the weighting schemes for keyword extraction for gene clustering. The result produced from TF-IDF weighted keywords outperformed those produced from normalized z-score weighted keywords. This result, therefore further justifies the effectiveness of using TF-IDF for weighing keywords in documents collection.

Finally, in Hany et al. (2007), XML technology was integrated with Information Retrieval scheme (TFIDF) and use Data Mining technique for association rules discovery. However, our work is focused on selecting the most discriminative keywords for use in association rules generation by combining syntactic and semantic relevance into the Information Retrieval Scheme which is also integrated with XML technology.

## TEXT MINING SYSTEM ARCHITECTURE

The proposed text mining system architecture is described in Figure 1. The system automatically discovers association rules from textual documents. The main contributions of the technique are that, in selecting the most discriminative keywords for use in association

*Figure 1. Improved text mining system architecture*


TEXT PREPROCESSING PHASE
Stopword list
Proxen Antelope-wordNet
Incoming Textual Documents
Filtration
Stemming
Index documents by using the weighting scheme TF-IDF for all keywords in all documents
Transform Documents to XML format
KNOWLEDGE DISTILATION PHASE
Apply the modified GARW algorithm on the indexed documents to generate all keyword sets whose support is greater the user specified minimum support (min support)
Generate all Association Rules that satisfied a user minimum confidence (min confidence)
Visualize Association Rules in textual format

rules generation, the system combines syntactic and semantic relevance into its Information Retrieval Scheme which is integrated with XML technology. It combines the above with Data Mining techniques for association rules extraction. The system ignores the order in which the words occur, but instead focuses on the words, their statistical distributions and their semantic relevance. The system begins with selecting collections of documents gotten from questionnaires. The Text Mining system consists of three phases: Text Preprocessing phase (transformation, filtration, stemming and indexing of the documents), Association Rule Mining (ARM) phase and Visualization phase (visualization of results).

- **The Text Preprocessing Phase:** This phase is aimed at optimizing the performance of the knowledge mining phase. It consists of text filtration, stemming, indexing and refinement of the extracted keywords based on a combination of both syntactic relevance (TF-IDF) weight and semantic similarity threshold. After this, the document is transformed to an XML format.
- **Filtration:** A word is selected as a keyword if it does not appear in a pre-defined stop-

words list. The stop-words list consists of articles, pronouns, determinants, prepositions and conjunctions, common adverbs and non-informative verbs.

- **Stemming:** After the filtration process the system does word stemming, a process that removes a word's prefixes and suffixes (such as unifying both infection and infections to infection). Stemming is done by unifying word based on their dictionary meaning using the WordNet lexical database. Word Net is referenced through Proxem Antelope (http://www.proxem.com/Default.aspx?tabid=55, which is a framework that makes the development of Natural Language Processing software easy to use. Proxem Antelope is designed to load WordNet files into the memory so as to make searches amazingly fast.
- **Indexing and Refinement:** In this stage, the weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency) is combined with semantic relevance weight to give a combined relevance weight as stated below.

The TF-IDF is used to assign higher weights to syntactically distinguished terms in a document, and it is the most widely used weighting scheme which is defined as (Feldman et al., 1996; Hany et al., 2007 ; Teng-Kai et al., 2010; Fang-Yie et al., 2010):

$$w(i,j) = tfidf(d_i, t_j) = \left\{ \begin{array}{ccc} Nd_{i,}t_j * \log_2 \dfrac{|c|}{Nt_j} & if & Nd_i, t_j \geq 1 \\ 0 & if & Nd_i, t_j = 0 \end{array} \right\} \quad (1)$$

- w(i,j) is known as the weighting scheme and could be greater than 0.
- $Nd_i,t_j$ is the number of times the term $t_j$ occurs in the document $d_i$.
- $Nt_j$ is the number of documents in the collection C in which the term $t_j$ occurs at least once.
- |C| is the number of documents in the collection C.

In general, this weighting scheme includes the intuitive presumption that the more often a term occurs in a document, the more it is representative of the document (term frequency) and the more the documents the term occurs in, the less discriminating it is (inverse document frequency). The system sorts the keywords based on their scores and selects them based on the given weight chosen as threshold.

The semantic relevance is gotten by exploiting the degree of polysemy of terms i.e. we want to weigh the semantic relevance of a term with respect to a notion of semantic rarity, in such a way that the higher the number of meanings of the term, the lower its rarity. Thus, its relevance. Since we assume that the terms in the XML data have been reduced to their stems, then, the semantic relevance is computed based on the polysemy of its variant terms that originally appear in the text (Andrea et al., 2010).

$$s - rarity(w) = \frac{1}{|\, o - terms(w) \,|} \left[ \sum_{w_j e o - terms(w)} \ln\left(\frac{MAX - POLYSEMY + 1}{|\, sense(w_j) \,| + 1}\right. \right] \quad (2)$$

- **T:** Collection of XML tree tuples i.e. a set of transactions
- **W:** Index term i.e we pick each term one by one
- **o-terms(w):** Set of original terms in T having w as the common stem
- **| o-terms(w)|:** Number of terms in T that their stem is w i.e. the particular term in question.
- **| senses($w_j$)|:** The number of meanings of $w_j$
- **MAX-POLYSEMY:** A constant denoting the number of meanings of the most polysenous term in the reference lexical knowledge base.
  - Note: The MAX-POLYSEMY depends on the part of speech of the selected terms e.g. its 32 for nouns in WordNet 2.0.

Combination of syntactic and semantic relevance to get the relevance weight of each term i.e. $w_j$ (Andrea et al., 2010).

$$relevance(w_j, u_i) = \frac{1 + s - rarity(w_j)}{|T(u_i)|} \sum_{\tau e T(u_i)} ttf.itf(w_j, u_i /\tau) \quad (3)$$

- **relevance($w_j$, $u_i$):** Stores the reference value of term $w_j$ in TCU
- **s-rarity($w_j$):** Gotten from semantic relevance

$$\sum_{\tau e T(u_i)} ttf.itf(w_j, u_i /\tau) - \text{the TF} - \text{IDF weight}$$

- **T($u_i$)**: Total number of TCUs or transactions.
- **Transformation:** The text mining system accepts documents format in text (.txt), performs filtration, stemming and indexes them before converting to XML format amenable for further processing.
- **The Knowledge Distillation Phase:** Knowledge is distilled using the GARW (Generating Association Rules based on Weighting scheme) algorithm described below (Hany et al., 2007):

## Generating Association Rules Based on Weighting Scheme (GARW) Algorithm

Given a set of terms

A = {$w_1$, $w_2$, …………$w_n$} (4)
A set of indexed documents D
={$d_1$,$d_2$,……………..$d_n$} (5)

- $d_1$……$d_n$ are indexed documents that contains keywords.
- Those keywords are also members of A i.e. the general database of keywords.

## Association Rule

Association rule is one of the most important techniques in Data Mining. The problem of association rule mining deals with how to discover association rules that have support and confidence greater than the user-specified minimum support and minimum confidence. It is intended to capture dependency among items in the database.

The support of an item set is the fraction of transactions in the database that contain all the items in the database

$$Support(w_i w_j) = \frac{\text{support count of } W_i, W_j}{\text{Total number of documents}} \quad (6)$$

The confidence of rule a (association rule) $W_i \rightarrow W_j$ can be defined as the proportion of those transactions containing $W_i$ that also contain $W_j$.

$$confidence(w_i / w_j) = \frac{\sup port(w_i w_j)}{\sup port(w_i)} \quad (7)$$

The algorithm for generating association rules based on the weighting scheme is given as follows:

1. Scan the file that contains all the keywords that satisfy the threshold weight value and their frequency in each document.
2. Let N denote the number of top keywords that satisfy the threshold weight value.
3. Store the top N keywords in index file along with their frequencies in all documents, their weight values TF-IDF and documents ID in the following format: <doc-id><keyword>< keyword frequency><TF-IDF>
4. Scan the indexed file and find all keywords that satisfy the threshold minimum support. These keywords are called large frequency1-keywordSet $L_1$.
5. When K is greater than 2, (Note K is a keyword set having k-keywords sets). The candidate keywords $C_k$ of size K are generated from large frequent (k-1) keywords sets, $L_{k-1}$ that is generated in the last step.
6. Scan the index file, and compute the frequency of candidate keyword sets $C_k$ that is generated in step 4.
7. Compare the frequencies of candidate keywords sets with minimum support.

8. Large frequent keyword sets $L_k$, which satisfy the minimum in support, is found from step 7 above.
9. For each frequent keyword set, find all the association that satisfies the threshold minimum confidence.

- **Rule Post Processing:** The generated rules are refined by using parameters such as the support and confidence which in this case we already been included in the GARW algorithms above. One particular aspect of rule mining in text is that often a high support means the rule is too obvious and thus less interesting. Another technique that was used to remove unwanted rules is to specify stop rules i.e. rules that are common and can be removed automatically. Association rules are easy to understand and to interpret for an analyst or a normal user. However, it should be mentioned that the association rule extraction is of exponential growth and a very large number of rules can be produced.
- **Rule Visualization Phase:** Even though association rules extracted from the above phases can be reviewed in textual format or tables, or in graphical format, in this work the system is designed to visualize the extracted association rules in textual format or tables.

## IMPLEMENTATION

The system was implemented using C# programming language and Visual Studio.Net 2005 as the programming environment. It loads the documents, receives three thresholds from the user, runs the program and displays the generated association rules.

- **Data Description:** The primary means of gathering data in our field of application, which is CRM is through the use of questionnaires. A questionnaire was therefore designed and administered to 2,215 respondents out of which 1,518 were returned valid. These questionnaires were designed with the goal of retrieving CRM information from mobile phone users towards effective customer relationship management in the mobile phone manufacturing industry. This questionnaire was justified through a pilot study and meeting with experts in CRM field. The questionnaire contained both structured and unstructured part. But only the unstructured part was used for this experiment. The following are the samples of questions asked in order to gather the unstructured data;

    - What do you like most about your mobile phone?
    - What do you dislike most about WAP?
    - Share your best mobile phones experience.
    - Why did you decide to purchase that particular brand of mobile phone?
    - What improvements would you like to see, if any on your mobile phone.
    - What type of problem do you usually encounter while using your mobile phone?

## Description of Extracted Association Rules:

### *Argumentation of the Thresholds*

In text mining in general, a very large number of association rules are found. So the measures like support and confidence are important when creating keyword sets and selecting the final rules. However, the problem is that we may find the important keywords which have frequently appeared recently but not discovered because the height of support and confidence threshold values. In order to have a fair representation of the important keywords in the corpus to be mined, we selected a combined relevance weight of 30%. This helped us to find informative keywords to extract rules from. Furthermore, a low threshold support of 5% was used so as to extract important keywords (such as durability, brand) that would not have appeared if we chose high support value, and these keywords happen to be very informative regarding customer relationship management as regards mobile phones. Lastly, we chose higher threshold confidence value 70% to make sure that the final rules gotten from the system are the most interesting ones.

## Discussion and Interpretation of the Extracted Rules

Some of the association rules that describe the relations between keywords in the documents are presented below. The rules give information on some interesting patterns that can be used for customer relationship management in the Nigerian mobile phone industry. In the first data grid, the first column represents rules generated and the second column represents the confidence.

Samples of the generated rules are seen in Figure 2 and interpreted for customer relationship management as follows:

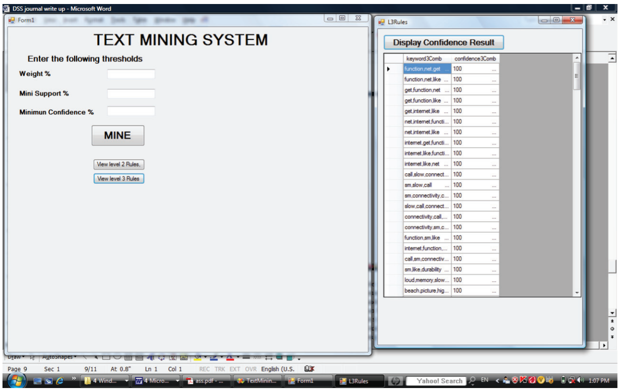
- The rule **plenty, people -> text:** gives the inference that the SMS facility on the mobile phone is used by plenty people. Therefore any mobile phone producing company can look for means of making this feature more user friendly so as to attract more customer to their own brand of mobile phone.
- The rule **internet, screen -> problem** shows that there is a problem with the screen of mobile phones while browsing the internet.
- The rule **improvement, speed -> nokia** reveals that nokia phone users want an improvement in the speed of their mobiles phones.
- The rule **long, battery->nokia**, can also help to infer that the mobile phone users like the battery time of their phones to be long.

We design another system for extracting association rules from text by using the GARW algorithm. This system corresponds to our system in the following processes:

- Transformation of documents into XML format;
- Filtration and stemming of the transformed documents;
- Reduction of keywords using the TF-IDF weighing scheme.

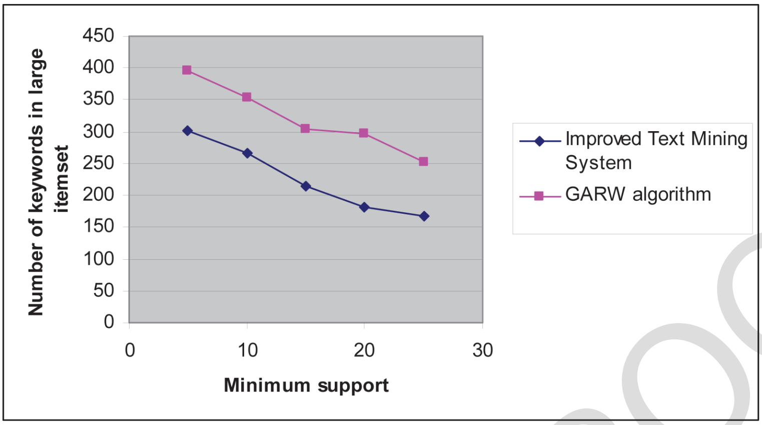
To assay the performance of the our system, we compared the large itemsets (first step of the association rule mining phase) generated from our system for different support thresholds with that of the one generated by the GARW algorithm (see Figure 3). This is done while keeping the value of the combined relevance weight threshold (for our system) and TF -IDF threshold (for the GARW system) constant at

*Figure 2. Association rules visualized in textual format*



| keyword3Comb | confidence3Comb |
|---|---|
| function,net,get ... | 100 ... |
| function,net,like ... | 100 ... |
| get,function,net ... | 100 ... |
| get,function,like ... | 100 ... |
| get,internet,like ... | 100 ... |
| net,internet,functi... | 100 ... |
| net,internet,like ... | 100 ... |
| internet,get,functi... | 100 ... |
| internet,like,functi... | 100 ... |
| internet,like,net ... | 100 ... |
| call,slow,connect... | 100 ... |
| sm,slow,call ... | 100 ... |
| sm,connectivity,c... | 100 ... |
| slow,call,connect... | 100 ... |
| connectivity,call,... | 100 ... |
| connectivity,sm,c... | 100 ... |
| function,sm,like ... | 100 ... |
| internet,function,... | 100 ... |
| call,sm,connectiv... | 100 ... |
| sm,like,durability ... | 100 ... |
| loud,memory,slow... | 100 ... |
| beach,picture,hig... | 100 ... |

*Figure 3. Improved text mining system vs GARW algorithm*



30% and the experiment was perform on the same corpus.

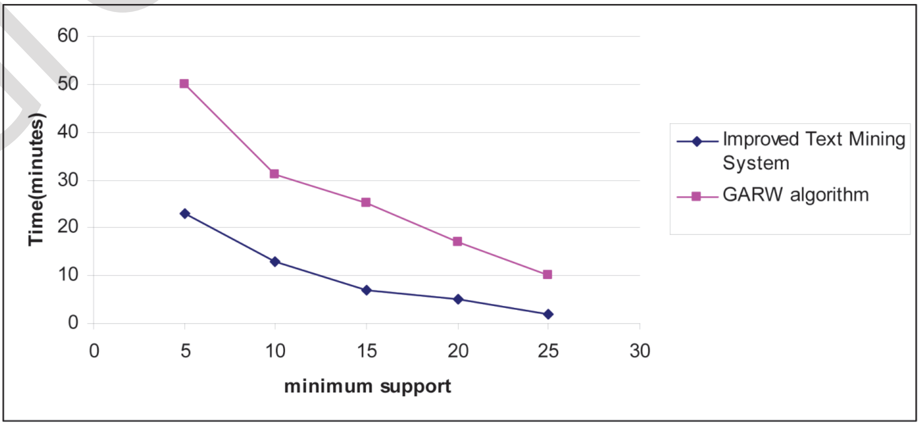
The experimental results displayed in the graph in Figure 3 reveals a reduction in the large itemset size generated from our system compared to the GARW algorithm. Also, the execution time of our system was compared with the GARW algorithm at a combined relevance weight threshold (for our system) and TF-IDF threshold (for the GARW system) of 30% for both systems and a confidence threshold of 70% for the same document corpus (100 documents was used form the purpose of this test) to reveal the results displayed in Figure 4.

It can be seen that our system always outperforms the GARW algorithm for all values of minimum support.

## CONCLUSION

The proposed approach is domain-independent, so it is flexible and can be applied on different domains without having to build a domain specific stemming dictionary. since identifying user requirements and understanding the user is a major part of contributing to the profit of the organization and this can be achieved through an effective customer relationship management.

*Figure 4. Graph of execution time against support*



The generated rules therefore, give pointers to various characteristics of customers of mobile phone manufacturing industry which will help in identifying, attracting, developing and maintaining successful customer relationships over time in order to increase retention of profitable customers.

Also, the results gotten from the experiments is traceable to the fact that, since the large itemset is responsible for the keyword combination stage (which accounts for most of the execution time) of the association rule mining algorithm, therefore the smaller the size of the large itemset, the more faster the algorithm execution time and the more semantically relevant the keywords in the large itemset, the more interesting the rules will be.

## FUTURE RESEARCH

We plan to extend our system by evaluating the resulting rule gotten of our association rule mining phase by using an evaluation techniques which reveals interestingness based on personalized knowledge discovery from text. That is, evaluate the novelty of discovered knowledge in the form of association rules by measuring the semantic distance between the antecedent and the consequent of a rule in the background knowledge. We also intend to visualize the extracted association rules in two –dimensional graphical representation.

## REFERENCES

Ah-Hwee, T. (1999). Text mining: The state of the art and the challenges. *In Proceedings of the PAKDD'99 Workshop on Knowledge Disocovery from Advanced Databases,* Beijing, China (pp. 65-70.)

Andrea, T., & Sergio, G. (2010). Semantic clustering of XML documents. *ACM Transactions on Information Systems, 28*(1). *DOI* = 10.1145/1658377.1658380 http://doi.acm.org/10.1145/1658377.1658380. Pp 3:1-3:55.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York, NY: ACM Press.

Berardi, M., Malerba, D., Marinelli, C., Leo, P., Loglisci, C., & Scioscia, G. (2005). *A text-mining application able to mine association rules from biomedical texts*. Retrieved from http://www.itb.cnr.it/bits2005/abstract/11.pdf

Chen, X., & Wu, Y. (2001). Personalized knowledge discovery: mining novel association rules from text. Retrieved from www.siam.org/meetings/sdm06/proceedings/067chenx.pdf

Fang-Yie, L., & Chih-Chieh, K. (2010). An automated term definition extraction system using the web corpus in the Chinese language. *Journal Of Information Science and Engineering*, *26*, 505–525.

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). *In Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining* (pp. 112–117).

Feldman, R., & Hirsh, H. (1996). Mining associations in text in the presence of background knowledge. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, WA.

Graham, H., John, S., & Nigel, P. (2004). *Marketing strategy and competitive positioning*. Pearson Education Limited.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Hany, M., Dietmar, R., Nabil, I., & Fawzy, T. (2007). A text mining technique using association rules extraction. *International Journal of Computational Intelligence*, *4*(1), 1304–2386.

Jan, P., & Peter, B. (1999). Text mining for documents annotation and ontology support. IST-1999-20364. In *Webocracy: web technologies supporting direct participation in democratic processes*. Retrieved from http://people.tuke.sk/jan.paralic/papers/BookChapter.pdf

Kernochan, W. (2006). *XQuery and XML data: DB2 helps manage the era of unstructured data.* Retrieved from http://searchdatamanagement.techtarget.com/tip/XQuery-and-XML-data-DB2-helps-manage-the-era-of-unstructured-data

Liu, Y., Navathe, S., Pivoshenko, A., Dasigi, A., Dingledine, R., & Ciliax, B. (2006). Text analysis of medline for discovering functional relationships among genes: evaluation of keyword extraction weighting schemes. *International Journal of Data Mining and Bioinformatics*, *1*(1). doi:10.1504/IJDMB.2006.009923 PMID:18402044.

Michal, L., Martin, S., Emil, G., Zoltan, B., & Ladislav, H. (2004). *Ontology based text annotation.* Retrieved from http://laclavik.net/publications/P626_ios_press.pdf

Nasukawa, T., & Nagano, T. (2001). Text analysis and knowledge mining system. *IBM Systems Journal, 40*(4), 967–984. ISSN:0018-8670.

Norvag, K., Eriksen, T., & Skgstad, K. (2006). *Mining association rules in temporal document collections*. Retrieved from http://www.idi.ntnu.no/~noervaag/papers/ISMIS2006.pdf

Rajman, M., & Besancon, R. (1997). Text mining: Natural language techniques and text mining applications. In *Proceedings of the 7th Working Conference on Database Semantics (DS-7)* (pp. 7-10). Chapan &Hall IFIP Proc. Series. Leysin, Switzerland.

Raymond, J. M., & Razvan, B. (2006). Mining knowledge from text using information extraction. *SIGKDD Explorations*, *7*(1), 3–10.

Raymond, J. M., & Un Yong, N. (2005). Text mining with information extraction. In *Proceedings of the 4th International MIDP Colloquium,* Van Schaik Pub., South Africa (pp. 141-160).

Shenzhi, L., Tianhao, W., & William, M. P. (2004). Distributed higher order association rule mining using information extraction from textual data. *SIGKDD Explorations*, *7*(1), 26–35.

Solomon, N., & Paul, G. (2003). Business intelligence. In *Proceedings of the Ninth Americas Conference on Information Systems* (pp. 3190-3199).

Strauss, J., El-Ansary, A., & Frost, R. (2006). *E-marketing* (International ed., Vol. 30, pp. 135–168). Pearson Prentice Hall.

Teng-Kai, F., & Chia-Hui, C. (2010). Exploring evolutionary technical trends from academic research papers. *Journal of Information Science and Engineering*, *26*, 97–117. http://www.proxem.com/Default.aspx?tabid=55.

*Fatudimu Ibukun Tolulope holds a BSc in Engineering Physics and MSc in Computer Science. She is currently a PhD student in the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. Her research interest is in the field of Data Mining. She is an Assistant Lecturer in the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. She enjoys reading and engages in creative arts.*

*C. K. Ayo holds a BSc, MSc, and PhD in Computer Science. His research interests include: mobile computing, Internet programming, e-business and government, and object oriented design and development. He is a member of the Nigerian Computer Society (NCS), and Computer Professional Registration Council of Nigeria (CPN). He is currently an Associate Professor of Computer Science and the Head of Computer and Information Sciences Department of Covenant University, Ota, Ogun state, Nigeria, Africa. Ayo is a member of a number of international research bodies such as the Centre for Business Information, Organization and Process Management (BIOPoM), University of Westminster. http://www.wmin.ac.uk/wbs/page-744; the Review Committee of the European Conference on E-Government, http://www.academic-conferences.org/eceg/; and the Editorial Board,* Journal of Information and communication Technology for Human Development.

*Charles Uwadia holds a BSc, MSc, and PhD in Computer Science. His research interests include Software Engineering. He is the present president of the Nigerian Computer Society (NCS), and Computer Professional Registration Council of Nigeria (CPN). He is currently a Professor of Computer Science in the University of Lagos, Nigeria, Africa.*

# International Journal of Knowledge-Based Organizations

*An official publication of the Information Resources Management Association*

## Mission

The mission of *International Journal of Knowledge-Based Organizations* (IJKBO) is to provide an international forum for organizational and governmental practitioners, researchers, information technology professionals, software developers, and vendors to exchange useful and innovative ideas within the field. This journal emphasizes the presentation and distribution of groundbreaking, original theories and concepts shaping future directions of research enabling business managers, policy makers, government officials, and decision makers to comprehend advanced techniques and new applications of information technology. IJKBO encourages exploration, exploitation, and evaluation of different principles, processes, technologies, techniques, methods, and models in sustainable knowledge ecosystems.

## Subscription Information

IJKBO is published quarterly: January-March; April-June; July-September; October-December by IGI Global. Full subscription information may be found at www.igi-global.com/ijkbo. The journal is available in print and electronic formats.

Institutions may also purchase a site license providing access to the full IGI Global journal collection featuring more than 100 topical journals in information/computer science and technology applied to business & public administration, engineering, education, medical & healthcare, and social science. For information visit www.igi-global.com/isj or contact IGI at eresources@igi-global.com.

## Copyright

**Correspondence and questions:**

**Editorial:**
John Wang
Editor-in-Chief
IJKBO
E-mail: journalswang@gmail.com

**Subscriber Info:**
IGI Global
Customer Service
701 E Chocolate Avenue
HersheyPA17033-1240,USA
Tel: 717/533-8845 x100
E-mail: cust@igi-global.com

*International Journal of Knowledge-Based Organizations* is listed or indexed in: Bacon's Media Directory; Cabell's Directories; DBLP; Google Scholar; INSPEC; JournalTOCs; MediaFinder; ProQuest Advanced Technologies & Aerospace Journals; ProQuest Computer Science Journals; ProQuest Illustrata: Technology; ProQuest SciTech Journals; ProQuest Technology Journals; The Standard Periodical Directory; Ulrich's Periodicals Directory