Discovering Temporal Bisociations for Linking Concepts over Time

Corrado Loglisci and Michelangelo Ceci

Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro"
via Orabona, 4 - 70125 Bari - Italy
{loglisci,ceci}@di.uniba.it

Abstract. Bisociations represent interesting relationships between seemingly unconnected concepts from two or more contexts. Most of the existing approaches that permit the discovery of bisociations from data rely on the assumption that contexts are static or considered as unchangeable domains. Actually, several real-world domains are intrinsically dynamic and can change over time. The same domain can change and can become completely different from what/how it was before: a dynamic domain observed at different time-points can present different representations and can be reasonably assimilated to a series of distinct static domains. In this work, we investigate the task of linking concepts from a dynamic domain through the discovery of bisociations which link concepts over time. This provides us with a means to unearth linkages which have not been discovered when observing the domain as static, but which may have developed over time, when considering the dynamic nature. We propose a computational solution which, assuming a time interval-based discretization of the domain, explores the spaces of association rules mined in the intervals and chains the rules on the basis of the concept generalization and information theory criteria. The application to the literature-based discovery shows how the method can re-discover known connections in biomedical terminology. Experiments and comparisons using alternative techniques highlight the additional peculiarities of this work.

1 Introduction

Data produced in real-world applications have become so complex, heterogeneous and time-varying that humans are overwhelmed when they attempt to conduct any analysis without technological help. Sophisticated software systems and, in particular, advanced Data Mining techniques are being continuously developed in order to support the analysis of such data and help the comprehension of the underlying phenomena. One of the Data Mining tasks well established but continuously studied is that of association discovery. Typically, associations are based on the notions of co-occurrence, inference of co-occurrence, correlation or similarity which often permit the extraction of useful and interesting connections, but which sometimes represent information already known to the user.

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part II, LNAI 6912, pp. 358–373, 2011.

Especially in the field of science, scientists need to create hypotheses worthy of being investigated and discover connections seemingly remote but supported by an intricate reasoning. Indeed, they have to handle large quantities of data of different natures, intrinsically complex and very often observed in dynamic processes whose structure, components and representation change over time. A part of this problem is elegantly addressed by the approaches which investigate the task of bisociation discovery [1],[13],[9]. By refining the original notion provided in [7], a bisociation is widely recognized as a link that connects concepts from two or more contexts, which are unconnected according to the specific view (very often corresponding to a subjective perspective) by which the contexts are defined. Contexts can be considered as distinct domains which collect a set of concepts, while the discovery of bisociations corresponds to an explorative process which crosses various domains and links concepts present in such domains.

To perform this discovery process two issues have to be addressed[1]: the representation or modeling of the domains and the strategy used to explore the modeled domains and to identify adequate concepts for the linkage. First, the existing approaches exploit a network-based representation which permits the aggregation of various domains (each of which associated to a sub-network) and relates concepts in the same domain and in different domains. The nodes are assigned to the concepts, while the edges express the relationships among the concepts directly observed in the domain or computationally derived, such as the relations of similarity, co-occurrence or probabilistic dependence. Second, interactive navigation techniques and graph analysis algorithms are used to explore the overall network and identify paths which, crossing several sub-networks (distinct domains), link nodes (or other sub-networks) which are far apart in the network and express valuable implicit relations.

Most of the approaches which implement this process rely on the assumption that domains are static, that is, disregard the dynamic nature of several real-world domains and solve only a part of the initial problem of the scientific discovery. Indeed, the network-based representation introduced above models a set of heterogeneous but unchangeable domains, while even a single domain can show changes over time. Indeed, a dynamic domain observed at different time-points can present different representations and can be reasonably assimilated in a series of distinct static domains. Hence, the necessity to investigate the problem of bisociation discovery arises when taking into account the temporal component and the intrinsic time-varying nature of some domains. This is also justified by the fact that temporal dynamics is attracting interest in the recent data mining literature, since it can play an important role in the comprehension of the evolution of domains under investigation. Among the most significant works, Bottcher et al[2] propose a paradigm based on the temporal dynamics to detect and quantify changes in time-varying models and patterns, while Kleinberg[6] investigates possible approaches to analyze stream-based data from a perspective which considers the temporal evolution of the information.

In this paper, we investigate the task of linking concepts from a dynamic domain through the discovery of bisociations which link concepts over time.

Bisociations permit the representation of linkages which may be unearthed only when considering the dynamic nature and which can not be discovered when considering the domain as static. The two issues (previously described) to perform the process of bisociation discovery are addressed in this work as follows. First, a time interval-based discretization is produced on the domain, so that two time intervals have two different representations of the same domain and therefore, they present somehow two different static domains. Searching for linkages in different domains, even if the data are modeled with the usual network-based representation, can raise computational problems, since the discovery process should evaluate possible links at the level of the original data. To overcome this issue we propose representing each static domain (corresponding to a time-interval) with an abstract description which permits us to focus on the main characteristics of the domain in that time-interval, hence the solution for the other issue. Second, an explorative process across the time-intervals performs the discovery of bisociations among concepts by chaining the abstract descriptions, which involve those concepts, on the basis of the concept generalization and information theory criteria. This allows us to avoid meaningless bisociations and to limit computational problems due to the exploration in the space of the abstract descriptions.

The paper is structured as follows. In the next section we report works related to ours and highlight some peculiarities of the current approach. In Section 3 we formalize the scientific problem studied in this work and in Section 4 we describe the computational solution we propose. The approach is tested with the application to literature-based discovery. Finally, conclusions are drawn.

2 Related Works and Contribution

Current research on bisociations focuses mainly on the discovery of unexpected links from heterogeneous domains by merging conceptual categories. In [9] the authors explore this problem in the analysis of microarray data by proposing a composite framework: the creation of a network-based representation with the integration of different biological repositories and ontologies, grouping of differentially expressed genes (concepts) with a subgroup discovery approach, and discovery of links among the genes contained in the groups. Links are discovered according to a probabilistic approach. A network derived from heterogeneous data sources is also realized in [1] where the authors implement a different discovery approach. In this approach nodes of the networks are assigned to annotated units of information (keywords, gene names), while the edges are weighted to express the degree of certainty and specificity of the relations in the domain between two nodes. Bisociations are discovered with a spreading activation algorithm which is able to extract subnetworks consisting of the most relevant nodes related to a specified set of initially activated nodes. In [13] the problem is explored in document collections, where the application of text-processing techniques permits to obtain a network: the nodes correspond to named entities annotated with a term-frequency vector while the edges are constructed on the basis of the co-occurrence of the entities in the documents. Bisociations are finally discovered by evaluating the similarity among nodes with vector-based similarity measures. A similar representation is used in [5], where a method combining frequent itemset mining and link analysis is proposed to identify chains of named entities and verbal forms (concepts) extracted from texts. A graph-structure is created by assigning frequent 2-itemsets (pairs of concepts) to a pair of nodes connected by an edge. Final chains of concepts are obtained by following walking paths with an interactive technique which uses statistical measures.

Finding links between seemingly unrelated concepts from a text is a research line started by the pioneering work of Swanson [14] and continued by several approaches of biomedical literature-based discovery. The blueprint of these methods is the A-B-C model [14] where two concepts A and C are given and the discovery process aims to identify the intermediate concept B. Typically, two disjointed literature sets are separately analyzed to mine connections like $A \Rightarrow B$, $B \Rightarrow C$, based on similarity, co-occurrence or correlation. The application of the transitive law would allow us to derive novel connections $A \Rightarrow C$. To obtain the connections $A \Rightarrow B$, $B \Rightarrow C$ two possible strategies can be identified in the literature: *closed discovery*, where the concepts A and C are provided by the user, and *open discovery* in which only A is given. Approaches working on the first strategy have focused mainly on the automatic tools to select the intermediate B concept, for instance in [4] connections are extracted in the form of association rules and the possible intermediate concepts are identified with the integration of domain ontologies. For the second strategy, particular attention is paid to the pruning techniques which eliminate meaningless connections $B \Rightarrow C$. For instance, in [10] the idea is that of considering terms B with respect to a term-frequency based measure, named *rarity*, while in [11] the authors propose knowledge-based heuristics to provide a ranking of the connections $B \Rightarrow C$.

Therefore, linking concepts over time seems to be a not yet investigated issue that would facilitate the discovery of connections between concepts only through a linking process over time. It is noteworthy that the problem here investigated is not different from the bisociation discovery seen in [1],[13],[9] where bisociations connect concepts from unconnected domains identified according to some view of data. In this work, we use the view inherently introduced by time. This means that each domain corresponds to a sort of snapshot of the dynamic domain. The representation of the data is another distinguishing aspect of the current approach from the others. Indeed, the domain is modeled with abstract representations in the form of lattice-based structures of multiple level association rules. Since rules denote statistical evidence, the usage of abstract descriptions justifies the robustness of the method by reducing the risk of false positive links.

3 Formal Definition of the Problem

Before formally defining the scientific problem of interest in this work, here we introduce some preliminary concepts. Let $O_D : \langle O_1, O_2 \ldots O_i \ldots O_n \rangle$ be a sequence of time-ordered observations on the concepts $\mathcal{C}$ of the domain D. For instance, in the domain of biomedical literature the set $\mathcal{C}$ would correspond to a set of biomedical named entities, while each observation O_i is assigned to a single paper with a specific publication date. Therefore, a subset of the named entities $\mathcal{C}$ is observed in a paper O_i.

Given a language $\mathcal{L}$ defined on the concepts $\mathcal{C}$, let $\mathcal{A}$ be a set of statements in $\mathcal{L}$ produced by applying an operator $\mathcal{M}$ to a subsequence $O_i \dots O_j$ of O_D ($i < j$). $\mathcal{M}$ provides abstract descriptions of subsequences $O_i \dots O_j$ ($i, j = 1 \dots n$). Each abstract description can be denoted with statistical parameters or certainty measures. By following the example above, $\mathcal{M}$ would generate frequent patterns $\mathcal{A}$ from the named entities in the set of papers $O_i \dots O_j$.

Let $\mathcal{T}$ be an operator which maps the set of time-stamps $\{t_1 \dots t_n\}$ of $\langle O_1, O_2 \dots O_i \dots O_n \rangle$ into τ, where $\tau : \{\tau_1, \dots \tau_n\}$ is a finite totally ordered set of time-points under the order relation denoted by "$\leq$". $\mathcal{T}$ permits to discretize time-stamps such that, given two time-stamps t_i, t_j for which $t_i < t_j$, also $\mathcal{T}(t_i) \leq \mathcal{T}(t_j)$. For instance, given three publication dates "April 20 2010", "May 10 2010", "May 10 2011": $\mathcal{T}$(April 20 2010)= $\mathcal{T}$(May 10 2010)= 2010, $\mathcal{T}$(May 10 2011)= 2011. Therefore, we can associate the sequence $O_D : \langle O_1, O_2 \dots O_i, O_{i+1}, \dots O_{n-1}, O_n \rangle$ of time-stamped observations with a sequence $\{\tau_1, \tau_2, \dots, \tau_i, \tau_{i+1}, \dots \tau_m\}$, $(\tau_1 < \tau_2 < \dots < \tau_m)$. For instance, given $O_D : \langle \mathit{April}\ 1\ 2008, \mathit{April}\ 2\ 2008 \dots \mathit{May}\ 1\ 2008, \mathit{May}\ 2\ 2008, \dots\ \mathit{April}\ 1\ 2009, \mathit{April}\ 2\ 2009, \dots \mathit{April}\ 1\ 2010, \mathit{April}\ 2\ 2010, \dots \mathit{April}\ 1\ 2011, \mathit{April}\ 2\ 2011, \rangle$, we can associate it with the sequence $\{2008, 2009, 2010, 2011\}$.

Definition 1. *Given X, Y concepts in $\mathcal{C}$, a temporal bisociation B is a sequence of abstract descriptions $A_1, A_2, \dots A_{m-1}$, $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2, \dots A_{m-1} \in \mathcal{A}_{m-1}$, where $\mathcal{A}_i$ is obtained from the observations included in $[\tau_i; \tau_{i+1}]$. A_1, A_{m-1} involve the target concepts X, Y respectively.*

Informally, Definition 1 states that, given a sequence of time-intervals $[\tau_1; \tau_2]$, $[\tau_2; \tau_3]$,... $[\tau_{m-1}; \tau_m]$, the abstract descriptions $\mathcal{A}_1, \mathcal{A}_2 \dots \mathcal{A}_{m-1}$ can be derived from them. The sequence $A_1, A_2, \dots A_{m-1}$ reports the bisociation from X to Y. Without loss of generality, in this work abstract descriptions are computed in the form of association rules so, for instance, the chain of rules $X \Rightarrow \mathbf{W}$, $\mathbf{W} \Rightarrow \mathbf{J}$, $\mathbf{J} \Rightarrow \mathbf{Z}$, $\mathbf{Z} \Rightarrow Y$ (**W,J,Z** sets of intermediate concepts) stands for a bisociation linking X to Y ($X \Rightarrow \mathbf{W}$, $\mathbf{W} \Rightarrow \mathbf{J}$, $\mathbf{J} \Rightarrow \mathbf{Z}$, $\mathbf{Z} \Rightarrow Y$ mined from the observations associated to four distinct consecutive time-intervals).

Considering the notions introduced so far, the problem of discovering temporal bisociations can be divided into two sub-problems:

1. *Given*: $O_D : \langle O_1, O_2 \dots O_i \dots O_n \rangle$, $\mathcal{T}$ such that the width of each time-interval $[\tau_r; \tau_{r+1}]$ is greater than or equal to a user-defined threshold $\Delta_{\mathcal{T}}$, a certainty measure C, *Find*: a set $R_{\mathcal{A}} : \{\mathcal{A}_1, \mathcal{A}_2, \dots \mathcal{A}_{m-1}\}$ of abstract representations satisfying the certainty measure C.
2. *Given*: two concepts $X, Y \in \mathcal{C}$, the set $R_{\mathcal{A}}$, a minimum number $\eta_{\mathcal{T}}$ of time-intervals to be crossed, a certainty measure M, *Find*: a collection $\mathcal{B}$ of temporal bisociations $A_1, A_2, \dots A_{m-1}$ meeting the certainty measure M with $(m-1) \geq \eta_{\mathcal{T}}$.

A computational solution to these sub-problems is described in the following.

4 Discovering Temporal Bisociations

We should remember that solving the two sub-problems previously formalized means addressing respectively the two issues introduced in the Section 1 when discovering bisociations. So, the approach which finds the set $R_{\mathcal{A}}$ actually permits us to define a representation of the domains. While, the approach which uses the set $R_{\mathcal{A}}$ to find the collection $\mathcal{B}$ integrates a domain exploration strategy in order to identify bisociations. The computational solution comprises a preliminary step aiming to exclude the trivial connections between X and Y, namely it checks that the two concepts are not already directly connected or that there is no obvious evidence which connects them.

4.1 Check for Direct Connections

Direct connections among the concepts X and Y can be expressed by either similarities or co-occurrences. In this work, we follow the second way, since the first one would require the usage of (dis)similarity measures, semantics and ontologies which can be cumbersome and computationally expensive in many applications. The check is performed by controlling the absence of statistical evidence of the connections both at the level of the static domains (each time-interval) and at the level of the dynamic domain (cross time-intervals). The technique to determine statistical evidence used in this work is that of association rule mining: rules which present the concepts on either the antecedent side or on the consequent side denote reasonably direct connections between the concepts, since X and Y co-occur in the set of supporting observations O_i. To perform this preliminary step we exploit the algorithm proposed in [8] which enables the discovery of non-redundant association rules. More precisely (Algorithm 1), the algorithm is first applied to the complete set of observations O_D (lines 4-8) and then separately to each partition $P \in \mathcal{P}$ produced by applying the operator $\mathcal{T}$ to O_D (lines 9-16). The width of each partition (time-interval) is forced to be bigger than $\Delta_{\mathcal{T}}$. A description of the algorithm is reported in the next section.

4.2 Generation of Abstract Descriptions

Once possible statistical evidences have been excluded, each of the partitions $\mathcal{P}$ of the observations O_D, produced by the application of $\mathcal{T}$, is represented as abstract representations in the form of association rules (ARs). These reflect the statistical regularities in the data of that partition $P \in \mathcal{P}$ and move the discovery of bisociations at upward to higher abstraction level resulting in a reduction of the risk of false positive links. In these terms, the certainty measure C in the formulated problem (Section 3) consists of the usual statistical parameters *support* and $confidence$ which denote the rules, so the resulting abstract descriptions are ARs, which meet the minimum thresholds of support and confidence. By following this idea, we integrate into the process of AR mining ontologies of the dynamic domain, which permit us to annotate or abstract the intermediate concepts of the links. Exploiting domain ontologies (or more generally, background

Algorithm 1. Check of direct connections.

```
input: $O_D, X, Y, ARM, minSup, minConf, \mathcal{T}, \Delta_{\mathcal{T}}$          output: $CHECK$
// $ARM$ algorithm of association rules mining
$CHECK := false$
$AR \leftarrow ARM(O_D)$ // $AR$ association rules mined from the data $O_D$
for all $R \in AR$ do
  if $X \in R$ and $Y \in R$ then
    $CHECK := true$
  end if
end for
$\mathcal{P} := partitioning(O_D, \mathcal{T}, \Delta_{\mathcal{T}})$
for all $P \in \mathcal{P}$ do
  $AR \leftarrow ARM(P)$ // $AR$ association rules mined from the data $P$
  for all $R \in AR$ do
    if $X \in R$ and $Y \in R$ then
      $CHECK := true$
    end if
  end for
end for
```

knowledge) on the concepts is not actually a novelty in bisociation discovery: in [9] the authors use biological ontologies to annotate gene sets, while, in this work, we resort to background $is - a$ hierarchies, which generalize the concepts and which exploit the is-a relationships among the occurring concepts to strengthen the reliability of links. The process of AR mining is performed by means of the algorithm proposed in [8] which enables the discovery of non-redundant ARs at several hierarchical levels from data represented in the quite simple form of attribute-value pairs. The possibility of pruning redundancies of that algorithm here turns out to be an important peculiarity which makes the resulting abstract descriptions compact and without superfluous information. We report a brief description of the algorithm of ARs mining in the following.

The algorithm is composed of two steps which permit respectively to i) generate the set of closed frequent itemsets (in this work, sets of concepts) at the different hierarchical levels whose support exceeds the minimum threshold and ii) discover from these itemsets ARs at the different hierarchical levels (multiple level ARs) whose confidence exceeds the minimum threshold. The first step implements the notion of *closed* itemsets when the items are hierarchically organized. Mining the closed itemsets from a set D of observations means mining the maximal elements of the equivalence classes of the all itemsets derived from D[1]. Hence, an itemset Y=$\langle$ y$_1$, y$_2$,. . . , y$_j$,. . . ,y$_h$$\rangle$ is closed iff no supersets of Y is supported by the same set of observations of Y, and therefore, by the same support of Y. When itemsets can be organized according to an is-a hierarchy H, the concept of closed itemset has to be extended to that of **multiple level closed itemset** to remove redundant (according to H) multiple level itemsets.

The algorithm proceeds by scanning the hierarchy in top-down mode while, at each level, it generates a set of multiple-level closed itemsets with increasing length. The length is the number of items present in an itemset. The second step

[1] Two itemsets belong to the same equivalence class when they cover the same observations.

extends the notion of *minimality* to the ARs derived by the itemsets produced in the first step. Formally speaking:

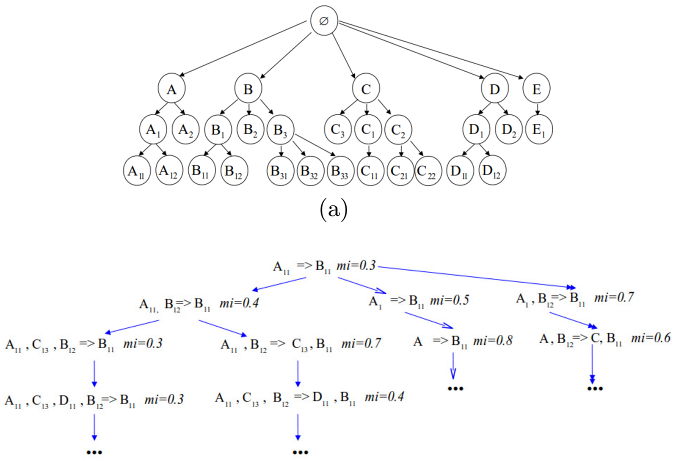
Definition 2. *An association rule $R_1 : A_1 \Rightarrow C_1$ is* ***minimal*** *iff $\nexists R_2 : A_2 \Rightarrow C_2$ with identical support and confidence of R_1, for which $A_2 \subseteq A_1$, $C_1 \subseteq C_2$.*

A interpretation of this definition for this work is that the minimal rules convey additional inferential information by means of the inclusion relationships of the antecedents and consequents of the rules. Indeed, if we consider $R_1 : A \Rightarrow B, C$, $R_2 : A, B \Rightarrow C$, with identical support and confidence, the antecedent of R_1 is included in the antecedent of R_2 while the consequent of R_1 includes the consequent of R_2. Hence R_1 is minimal with respect to R_2 and gives more information on the consequent side than R_2. R_2 is considered redundant.

4.3 Linking Concepts over Time

Once the ARs are discovered in each partition P on O_D, they are organized in a lattice-based structure ($R_{\mathcal{A}} : \{\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_{m-1}\}$): the nodes of the lattice represent ARs, while the edges represent relationships between the ARs. A finite sequence of edges which relate two ARs is called a *path*. Three types of path originate from the root of a lattice: a) paths for the generalization of the concepts contained in the root, b) paths for the extension of the rule at the root with larger rules, c) paths for the generalization of the concepts contained in the root with larger rules. In Figure 1b, $A_{11}, B_{12} \Rightarrow B_{11}$ is a node of the paths of type a (arrow), $A_1 \Rightarrow B_{11}$ is a node of the paths of type b (thick arrow), according to the hierarchy in Figure 1a, while $A_1, B_{12} \Rightarrow B_{11}$ is a node of the paths of type c (double arrow) according with the same hierarchy. In these three cases the rules are positioned in the lattice by child-father relationships of the hierarchy H and increasing length: that is, the rules at level $k+1$ contain father concepts of the concepts contained in the rules at level k and present one more concept than the rules at the level k. This permits an early evaluation of the rules which contain a low number of concepts among those occurring in the observations. Moreover, integrating heuristics on the paths and the organization of the nodes permits us to conduct an informed search in the lattices thus reducing the overall computational cost.

The value of a certainty measure M is associated to each rule: in this work mutual information plays the role of M introduced in the problem formulation (Section 3). Mutual information (mi) is one of the quantitative measures which can denote a rule and it has the peculiarity to express the mutual dependence between the antecedent and the consequent of a rule. What is more interesting is that it represents the ratio of the actual probability of two concepts to be related to the probability of two concepts to be unrelated. In this work, we prefer mi to other typical parameters, such as confidence. Actually, the confidence is not an appropriate measure of correlation strength between concepts since it leads to select common concepts in the consequents and rare concepts in the antecedents. The consequents in these cases rarely add much meaning to the final link. Differently, mutual information emphasizes relatively rare concepts that generally



(a)

(b)

Fig. 1. Is-a hierarchy over the concepts and a portion of the lattice of multiple-level association rules produced from a partition with their mi values

occur together and mitigates the importance of common concepts, thus leading to the discovery of more interesting bisociations. Mutual information is computed as $log\frac{support(Antecedent,Consequent)}{support(Antecedent)*support(Consequent)}$. ARs with mutual information less than a user-defined minimum threshold σ_{mi} are not considered.

The discovery of temporal bisociations is performed through an explorative process which crosses the lattices of the time-intervals and chains the included ARs by considering two different, complementary directions. The semantics associated to the concepts, expressed in the form of concept generalization and the information theory measure, in form of the mutual information are associated to each AR. Chaining is the basic operation which produces the link between two ARs: R_1, R_2 discovered in two consecutive time-intervals. Links are produced when the antecedent of R_2 contains either concepts present in the consequent of R_1 or more general concepts than those present in the consequent of R_1. A sequence of chained ARs constitutes a temporal bisociation: final bisociations have to include a number of ARs greater than or equal to a user-defined threshold $\eta_{\mathcal{T}}$. In other words, we are interested in temporal bisociations which have been developed over at least $\eta_{\mathcal{T}}$ consecutive time-intervals and which therefore involve at least $\eta_{\mathcal{T}}$ concepts. In particular, the root of the first lattice involved in a bisociation ($\mathcal{A}_1$) contains a rule (A_1), whose antecedent presents only the target concepts X, while the last lattice involved in the same bisociation ($\mathcal{A}_{m-1}$) contains a rule (A_1), whose consequent presents only the target concept Y. The total number of lattices to be explored has to be greater than or equal to $\eta_{\mathcal{T}}$: bisociations discovered from non-consecutive lattices or which do not cover this time span will be not considered.

The explorative process integrates a depth-first search by visiting the paths in this order: type a, b, c. In each lattice, the exploration visits the nodes by starting from the root: if the value of mi of the current node exceeds σ_{mi}, then the exploration in that lattice is completed and the antecedent of the rule in the current node is used as a "bridge" to explore the lattice of the next time-interval to link one of the contained rules. Otherwise, the search continues downward level-by-level by considering nodes of the same type of path until it reaches the leaf nodes. Then, it proceeds by backtracking and continues to explore paths of the same type or, when the paths of the same type have been completed, it goes back to the root and continues on paths of another type.

The linking process associated to the lattice of the next time-interval searches for a rule suitable for the linkage according to the following modalities: 1) rules of length two (one concept in the antecedent-one concept in the consequent), whose antecedents contain only the consequent of the final rule of the previous lattice; 2) rules of length greater than two, whose antecedents contain also the consequent of the final rule of the previous lattice; 3)rules of length two, whose antecedents contain only one concept which generalizes (according to the hierarchy H) the consequent of the final rule of the previous lattice; 4) rules of length greater than two, whose antecedents contain also a concept which generalizes the consequent of the final rule of the previous lattice. When several rules are identified as possible roots of the lattice to be explored, then the rule with higher mi value is selected. The discovery process continues by combining the exploration in each lattice (previously described) and the linking technique between lattices of the consecutive time-intervals up to the last lattice which completes the bisociation with a rule, whose consequent will be the target concept Y.

A trace of the discovery process is reported in Figure 3. Consider the time-intervals [1990;1992],[1992;1994], [1994;1996], σ_{mi}=0.5 and let A_{11} be the target object X. The process starts by searching in the lattice of [1990;1992] for the "entry" rule for the exploration, namely a rule whose antecedent is A_{11} (Figure 3a). Once identified, the exploration proceeds by evaluating mi of the rules with paths of type a: first the branch annotated with square 1, where no rule exceeding σ_{mi} is found, then the branch with square 2, where we have the same result. Subsequently, the path of type b (square 3) is explored, where the rule $A_{11} \Rightarrow B_{11}$ exceeds σ_{mi} (bold square 3): therefore the concept B_{11} becomes the bridge between the time-intervals [1990;1992],[1992;1994].

The exploration in [1992;1994] starts by searching for rules (with higher mi) whose antecedents contain B_{11} (Figure 3a). Once the "entry" rule has been identified (modality 3 above), the nodes of the branches with square 3, 4, 5 (types a, c) are evaluated, but none of them with success. Once the lattice has been completely explored, a new root is identified as the rule containing B_{11} on the antecedent (modality 3) and a new exploration in a new lattice of the same time-interval starts (Figure 3b). By following the same exploration strategy, the rule which continues the linkage is $A_{12}, C_{13}, B_1 \Rightarrow C, D_{11}$ (bold square 4). Hence, the list of concepts C, D_{11} becomes the bridge between the time-intervals [1992;1994],[1994;1996], as a further contribution to the bisociation. The

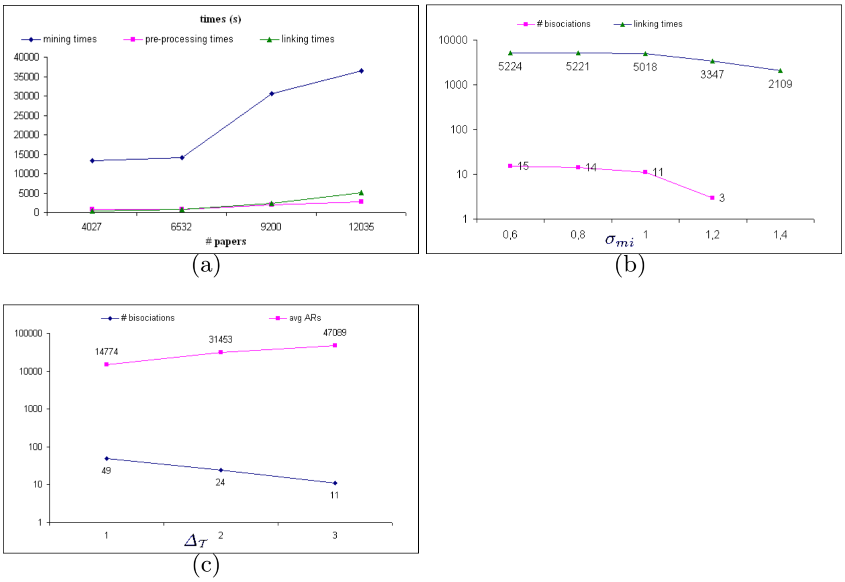


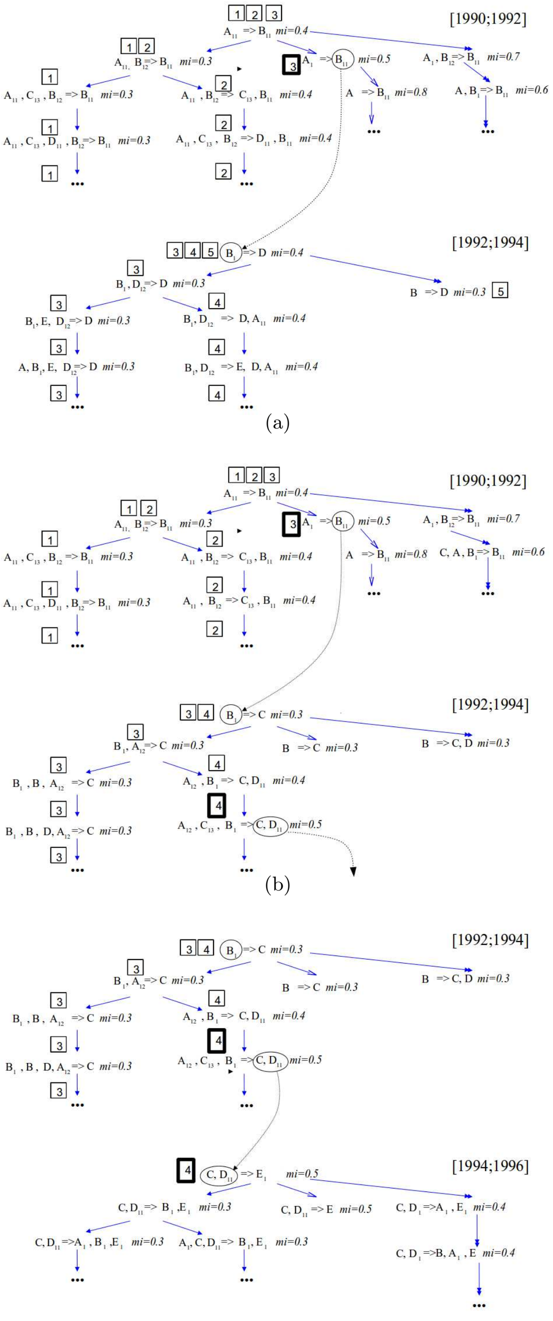
Fig. 2. Running times as a function of the number of papers published in [2000;2009] ($\Delta_{\mathcal{T}} = 1$) (a). Number of bisociations as a function of σ_{mi} (b) and $\Delta_{\mathcal{T}}$ (c).

exploration in [1994;1996] starts by searching for the rule (with higher mi), whose antecedents contain one of the possible permutations of C, D_{11} (Figure 3c). The identified root is the rule $C, D_{11} \Rightarrow E_1$ (modality 2), which provides also the bridge E_1 for the next interval.

Note that strategies to improve the exploration of the lattices may turn out to be ineffective in the case of informed searches. Moreover, pruning techniques would be inapplicable considering that, for the measure of mutual information, the *anti-monotonic* property does not hold for either rules with generalized concepts or rules with different length or with identical length.

5 Experiments on Biomedical Literature

One of the widely recognized dynamic domains is the scientific literature. It represents the typical source of information which researchers exploit for their studies and the typical means to disseminate their investigations. Publications may report studies on the same topic conducted one after another over time and this motivates our vision of the scientific literature as a dynamic domain. Literature is therefore a natural field to prove the viability of automatic tools for scientific discovery. For the current work, biomedical literature also represents the board on which to compare existing techniques with the one we are proposing. In this sense, these experiments aim at re-discovering known connections in biomedical terminology and highlighting potentialities offered by this work.



(a)

(b)

(c)

Fig. 3. Linking concepts over three time-intervals

Experimental Setup. Dealing with publications requires a necessary pre-processing step which permits the generation of the set of observations O_D. The original data set was composed of publications retrieved by the Pubmed search engine. In particular, since we tested our approach on the biomedical literature-based Swanson discoveries [14], the collection of original publications was generated from the result sets returned by Pubmed when, in December 2009, we submitted two distinct queries, namely "migraine", "magnesium deficiency". The two publication sets were obtained amounting to 5311 and 22223, respectively. The title, publication date and abstract sections were considered and pre-processed. Basic natural language processing techniques available in the GATE framework[2] were applied in order to identify biomedical named entities. This result was also obtained integrating controlled domain thesauri, such as MEsh Terms vocabulary[3], whose taxonomic organization allowed the production of the hierarchy H used for the generation of abstract descriptions (Section 4.2). The terminology available in the thesauri produces the set C of concepts.

The problem of the presence of synonyms was also addressed by integrating linguistic resources which permit the replacement of variant names of biomedical entities with their canonical names. The application of the operator $\mathcal{T}$ enables the discretization over time of the complete dynamics of the biomedical literature into partitions of publications published in time-intervals, based on the temporal dimension of the years. Each pre-processed paper corresponds to an observation O_i included in a time-interval and it can be interpreted as the investigation of the scientists at a specific time-stamp. The generation of the multiple-level ARs (abstract descriptions) was performed on sets of data, each of which is composed of the subset of pre-processed papers included in a time-interval. A further pre-processing was conducted by selecting the subset of concepts occurring in each paper whose TF-IDF measure[12] exceeded a user-defined minimum threshold.

Experiments were performed considering four different criteria, namely scalability, influence of the input parameters on the temporal bisociations patterns, information conveyed in the bisociations and comparison with existing solutions.

Scalability. Experiments on the performances in time were performed when increasing the threshold $\eta_{\mathcal{T}}$ and hence the number of papers, while the value of $\Delta_{\mathcal{T}}$ is 1 year (width of the time-intervals), $minSup$=0.3, $minConf$=0.7, σ_{mi}=1, minimum TF-IDF = 0.3. Collected running times consider the step of pre-processing, the AR mining algorithm and the discovery of temporal bisociations. In Figure 2a the results obtained considering the whole set of papers published in [2000;2009] are reported. They show that the computational cost is mainly due to the step of abstract description generation, which, however, returns a number of rules multiplied by a factor 3, while the number of papers increases of the same factor (from 4027 to 12035). This is also justified by the fact that the mining algorithm generates rules at different hierarchical levels, given that it integrates the hierarchy organizing the concepts. Indeed, when # papers is 12035 we have the highest number of ARs and the highest average of papers per time-interval

[2] `http://gate.ac.uk/family/`

[3] `http://www.nlm.nih.gov/mesh/`

(Table 1). The running times of the linking process are encouraging since it increases linearly with respect to the number of considered time-intervals (and therefore number lattices to be visited). This performance is due to the used heuristics which avoid a greedy exploration of the lattices.

Influence of Parameters. We tested the proposed computational solution when tuning the minimum threshold σ_{mi} and $\eta_{\mathcal{T}}$ ($minSup$=0.3, $minConf$=0.7). A initial consideration can be drawn from the results in Figure 2b (performed for the papers published in [1990;2008], $\eta_{\mathcal{T}}$=9, $\Delta_{\mathcal{T}}$=3 years), which empirically confirm the influence of the mutual information on the bisociations. It emerges that the most discriminative values of mi are basically included in the range (1;1.4], therefore, tuning σ_{mi} to values lower than 1 does not require additional computational cost and leads to the discovery of approximately the same set of bisociations. In fact, this is due to the replacement of variant names of concepts with canonical names, whose co-occurrences tend to strengthen their dependence. An interesting aspect is the generation of a maintainable set of bisociations which the user can easily investigate. This is due to fact that the setting requires the discovery of bisociations linking concepts over a relatively wide period of 27 years, namely 9 time-intervals, each of which covers 3 years. The experiments in Figure 2c (σ_{mi}=1, $\eta_{\mathcal{T}}$=9), on the influence of $\Delta_{\mathcal{T}}$, confirm the maintainability of the discovered bisociations, especially when compared to the number of rules (at the worst, 49 against 14774). Indeed, when $\Delta_{\mathcal{T}}$=1 the maximum number of lattices to be explored is generated, which generally leads to the strong increase of bisociations and to the reduction of the average number of ARs per lattice.

Comparison with Existing Techniques. The approach was compared with the existing systems $BITOLA$[3] and $ARROWSMITH$[15] focusing on the problem of literature-based discovery and on the same original data. These systems work in an interactive way, so comparing running times does not give any indications. By submitting the concepts X as "magnesium deficiency" and Z "migraine", $BITOLA$ discovers 2620 possible linking concepts Y able to form links with three concepts. For each pair (X, Y), (Y, Z), statistical parameters (e.g., frequency) are provided, although the huge set of intermediate concepts could be cumbersome for the user. On the other hand, our approach, by setting the target concepts X as "magnesium deficiency" and Y as "migraine", discovers only one temporal bisociation (even tuning σ_{mi} in [0,5;1]). By setting $minSup$=0.3, $minConf$=0.4, $\Delta_{\mathcal{T}}$= 1 year, $\eta_{\mathcal{T}}$=2 the following bisociation between "magnesium deficiency" and "migraine" is discovered in [1990; 1997] (the contribution of each intermediate concept is temporally collocated):

Table 1. Total and average abstract descriptions by increasing the number of papers

$[\eta_{\mathcal{T}}]$	$[\tau_1; \tau_m]$	# papers	# ARs	avg ARs
3	[2000;2003]	1342	15302	4027
5	[2000;2005]	2177	16143	6532
7	[2000;2007]	3066	22880	9200
9	[2000;2009]	4011	27908	12035

[1990;1991] Magnesium deficiency AND Anatomy ⇒ Metals, alkaline earth AND Metals [support=0.30, confidence=1.0, mi=1.19]

[1991;1992] Metals, alkaline earth AND Metals AND Diseases ⇒ Metals, light [support=0.33, confidence=0.97, mi=1.075]

[1992;1993] Chemicals and drugs AND Neurologic manifestations ⇒ Signs and symptoms [support=0.31, confidence=1.0, mi= 1.17]

[1993;1994] Central Nervous System Diseases AND Signs and Symptoms ⇒ Neurologic Manifestations [support=0.3, confidence=0.97, mi=1.178]

[1994;1995] Biological Sciences AND Neurologic Manifestations ⇒ Pathological Conditions, Signs and Symptoms [support=0.318, confidence=1.0, mi=1.144]

[1995;1997] Headache Disorders AND Pathological Conditions, Signs and Symptoms ⇒ Migraine [support=0.30, confidence=1.0, mi=1.185]

An identical comparison was performed with $ARROWSMITH$, which, however, requires more human intervention to carry out the process. The search for intermediate concepts from "magnesium deficiency" to "migraine" produces 598 possible links which can be further investigated with the support of the user, while the proposed solution requires less interaction.

We also evaluated the final bisociations by considering Swanson's linking terms [14] as gold standard. In his work, eleven hidden connections between "magnesium deficiency" to "migraine" were identified with the A-B-C model, whose intermediate concepts were: *Type A personality, vascular reactivity, calcium blockers, platelet activity, spreading depression, epilepsy, serotonin, inflammation, prostaglandins, substance P, brain hypoxia.* A subset of four terms, namely: epilepsy, serotonin, inflammation, prostaglandins, were contained in the controlled vocabulary and is-a hierarchy that we used, while the others were not recognized. Temporal bisociations discovered in [1989;1997] (period of investigation in [14]) involved effectively those four concepts and more general concepts (father concepts) than the former according to the hierarchy H.

Another consideration can be made from a basic aspect of the approach: the process of linking X to Y can produce different bisociations from linking Y to X over time, therefore, the temporal order is relevant in this work and permits us to determine the collocation over time of each contributing intermediate concept. For instance, the set of bisociations obtained from [1980;2009] (*minSup*=0.3, $minConf$ =0.4, σ_{mi}=0.8, $\Delta_{\mathcal{T}}$= 2 years, $\eta_{\mathcal{T}}$=2, X= "magnesium deficiency", Y="migraine") amounts to only one (not reported here due to lack of space) which involves thirteen concepts. If X="migraine", Y= "magnesium deficiency" one bisociation which involves seven intermediate concepts is discovered.

6 Conclusions

We have presented a novel approach to discover bisociations when considering the time-varying nature of the domains. Contrary to previous approaches, the discovery is performed on abstract descriptions of the domains which provide several advantages: focus on the main characteristics of the domains, prevention of computational cost due to the search in the original data and the reduction of the risk of false positive links. The linking process exploits two criteria, one based on semantics, the other one based on information theory. The application to the problem of literature-based discovery proves the reproducibility of the known results and the scalability of the approach. For future work, we plan to further improve the search in the lattices and extend experiments to other scenarios.

Acknowledgment. This work is supported by the research project ATENEO-2010: "Modelli e Metodi Computazionali per la Scoperta di Conoscenza in Dati Spazio-Temporali".

References

1. Berthold, M.R., Dill, F., Kötter, T., Thiel, K.: Supporting creativity: Towards associative discovery of new insights. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 14–25. Springer, Heidelberg (2008)
2. Böttcher, M., Höppner, F., Spiliopoulou, M.: On exploiting the power of time in data mining. SIGKDD Explorations 10(2), 3–11 (2008)
3. Hristovski, D., Peterlin, B., Mitchell, J.A., Humphrey, S.M.: Using literature-based discovery to identify disease candidate genes. I. J. Med. Inf. 74(2-4) (2005)
4. Hu, X., Zhang, X., Yoo, I., Wang, X., Feng, J.: Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule. Int. J. Intell. Syst. 25(2), 207–223 (2010)
5. Jin, W., Srihari, R.K., Ho, H.H., Wu, X.: Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. In: ICDM 2007, pp. 193–202 (2007)
6. Kleinberg, J.: Temporal dynamics of on-line information streams. Data Stream Management: Processing High-Speed Data Streams (2006)
7. Koestler, A.: The act of Creation. London Hutchinson (1964)
8. Loglisci, C., Malerba, D.: Mining multiple level non-redundant association rules through two-fold pruning of redundancies. In: Perner, P. (ed.) MLDM 2009. LNCS, vol. 5632, pp. 251–265. Springer, Heidelberg (2009)
9. Mozetic, I., Lavrac, N., Podpecan, V., Novak, P.K., Motain, H., Petek, M., Gruden, K., Toivonen, H., Kulovesi, K.: Bisociative knowledge discovery for microarray data analysis. In: 1st Int. Conf. on Computational Creativity, Lisbon, Portugal (2010)
10. Petric, I., Urbancic, T., Cestnik, B., Macedoni-Luksic, M.: Literature mining method rajolink for uncovering relations between biomedical concepts. Jour. of Biom. Inf. 42(2), 219–227 (2009)
11. Pratt, W., Yetisgen-Yildiz, M.: Litlinker: capturing connections across the biomedical literature. In: Gennari, J.H., Porter, B.W., Gil, Y. (eds.) K-CAP, pp. 105–112. ACM, New York (2003)
12. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company, New York (1984)
13. Segond, M., Borgelt, C.: Selecting the links in bisoNets generated from document collections. In: Cohen, P.R., Adams, N.M., Berthold, M.R. (eds.) IDA 2010. LNCS, vol. 6065, pp. 196–207. Springer, Heidelberg (2010)
14. Swanson, D.R.: Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine 31, 526–557 (1988)
15. Swanson, D.R., Smalheiser, N.R.: Implicit text linkages between medline records: Using arrowsmith as aid to scientific discovery. Library Trends 48(1) (1999)