PRIFYSGOL
ABERYSTWYTH
UNIVERSITY

Aberystwyth University

Order-of-magnitude based link analysis for false identity detection
Shen, Qiang; Boongoen, Tossapon

Published in:
Proceedings of the 23rd International Workshop on Qualitative Reasoning

Order-of-Magnitude Based Link Analysis for False Identity Detection

Tossapon Boongoen and Qiang Shen
Department of Computer Science, Aberystwyth University, UK

Abstract

Combating identity fraud is crucial and urgent as false identity has become the common denominator of all serious crime, including mafia trafficking and terrorism. Typical approaches to detecting the use of false identity rely on the similarity measure of textual and other content-based characteristics, which are usually not applicable in the case of deceptive and erroneous description. This barrier can be overcome through link information presented in communication behaviors, financial interactions and social networks. Quantitative link-based similarity measures have proven effective for identifying similar problems in the Internet and publication domains. However, these numerical methods only concentrate on link structures, and fail to achieve accurate and coherent interpretation of the information. Inspired by this observation, this paper presents a novel qualitative similarity measure that makes use of multiple link properties to refine the underlying similarity estimation process and consequently derive semantic-rich similarity descriptors. The approach is based on order-of-magnitude reasoning. Its applicability and performance are experimentally evaluated over a terrorism-related dataset, and further generalized with publication data.

Introduction

False identity has become the common denominator of all serious crime such as mafia trafficking, fraud and money laundering. Particularly in the UK, financial losses due to such cause are reported to be around 1.3 billion pounds each year (Wang *et al.* 2006). Holders of false identity are determined to avoid accountability and traces for law enforcement authority. In essence, such offence is intentionally committed with a view to perpetrating another crime from the most trivial to the most dreadful imaginable. Organized criminals make use of counterfeit identity to cover up illicit activities and illicitly gained capital. Especially in the case of terrorism, it is widely utilized to provide financial and logistical support to terrorist networks that have set up and encourage criminal activities to undermine civil society. Tracking and preventing terrorist activities undoubtedly requires authentic identification of criminals and terrorists who typically possess multiple fraud and deceptive names, addresses, telephone numbers and email accounts.

With present high-quality off-the-shelf equipment, it is almost effortless to obtain false identity documents. Conversely, it requires a great deal of time and experience to distinguish between genuine and forged copies. However, a successful detection can prevent the revolting consequence like that of shocking September-11 terrorist attacks. In particular to this tragedy, US authorities seriously failed to discover the use of false identities by nineteen terrorists, who were all able to enter the United States without any problem, in the very morning of the attacks. Most of them typically possess several dates of birth and multiple aliases (Boongoen & Shen 2008). For instance, *Mohamed Atta*, alleged ringleader of the September 11 attacks, has exploited several different aliases of *Mehan Atta, Mohammad El Amir, Muhammad Atta* and *Muhammad Al Amir Awad Al Sayad*. Identity verification and name variation detection systems (Wang *et al.* 2006) that rely solely on the inexact search of textual attributes may be effective in some cases. However, these methods would fail drastically to disclose unconventional truth of highly deceptive identity like that between *Osamabin Laden* and *The Prince* (Hsiung *et al.* 2005).

The aforementioned dilemma may be overcome through link analysis, which seeks to discover knowledge based on the relationships in data about people, places, things, and events. Intuitively, despite using distinct false identities, each terrorist normally exhibits unique relations with other entities involving in legitimate activities found in any open or modern society, making use of mobile phones, public transportation and financial systems. Link analysis techniques have proven effective for identity problems (Badia & Kantardzic 2005), (Hsiung *et al.* 2005) by exploiting link information instead of content-based information, which is typically unreliable due to intentional deception, translation and data-entry errors (Wang *et al.* 2005). Recently, link analysis is also employed by Argentine intelligence organizations to analyzing Iranian-Embassy telephone records in such a way to make a circumstantial case that the Iranian Embassy had been involved in the July 18, 1994, terror bombing of a Jewish community centre (Porter 2008).

Essentially, to justify the similarity between entities (e.g. names, publications and web pages) in a link network, many well-known algorithms like SimRank (Jeh & Widom 2002), PageSim (Lin, King, & Lyu 2006) and Connected-Triple (Klink *et al.* 2006) analogously concentrate only on the car-

dinality of joint neighbors to which they are directly linked, without taking into account the characteristics of a link itself. As such, the quality of the similarity evaluation may be enhanced by including uniqueness measure of links (Boongoen & Shen 2008) within the overlapping neighbor context. However, a definite precaution to combining multiple measures is the inaccuracy of quantitative descriptions, which are usually caused by a few link patterns with unduly high values. As a result, the measures of other patterns are very small and their interpretations become rather misleading.

In light of such shortcoming, this paper presents a novel link-based similarity measure that derives a qualitative similarity description from multiple link characteristics each expressed using the absolute order-of-magnitude model (Piera 1995). In essence, these properties are perceived at different precision levels, and hence being gauged in accordance to distinct orders of magnitude spaces. With different sets of measurement labels (i.e. landmarks), these scales differ by at least one qualitatively important order of magnitude. Particularly, a semi-supervised method is introduced to select data-driven landmarks, which are more reliable than those human-directed ones. In order to combine measures of multiple link properties, the homogenization of such references (Agell, Rovira, & Ansotegui 2000) is required to realize the ultimate similarity description, where relevance of properties is proficiently blended within the aggregation process.

The rest of this paper is organized as follows. Section 2 introduces the absolute order-of-magnitude model upon which the present research is developed. Following that, Section 3 describes link properties and order-of-magnitude based similarity evaluation. Section 4 presents the semi-supervised method for designing landmarks, which is data-driven and more robust than the human-directed counterpart. The experimental evaluation of this qualitative link-based similarity measure to detecting the use of false identity is detailed in Section 5. The paper is concluded in Section 6, with the perspective of further work.

Absolute Order of Magnitude Model

The absolute order of magnitude (AOM) model (Piera 1995) operates on a finite set of ordered labels or qualitative descriptors achieved via a partition of the real number line $\mathcal{R}$. Each element of the partition represents a basic qualitative class to which a label is associated. The number of labels selected to express each variable of a real problem is subject to both the characteristics and the precision level required to support comprehension and communication. In practice, multiple label sets with dissimilar granularities are typically utilized to define domain attributes qualitatively.

Despite the intuition that the number of labels is not fixed, the most conventional partitions are symmetric. That is, the partition of the underlying domain typically has n positive and n negative labels, which is formally represented by $OM(n)$, and referred to as the AOM model of granularity n. The real-line partition into $2n + 1$ labels is dictated by the set of $2n - 1$ landmarks. In essence, landmarks are domain dependent and determined by either subjective justification of human experts or learning from data. For instance, the $OM(3)$ model is built on the following set of landmarks:

$\{-\beta, -\alpha, 0, \alpha, \beta\}$. Figure 1 illustrates the resulting partition into seven qualitatively distinct order-of-magnitude labels, which are the most commonly used: Negative Large (NL), Negative Medium (NM), Negative Small (NS), Zero (0), Positive Small (PS), Positive Medium (PM) and Positive Large (PL) (Olmo *et al.* 2007).
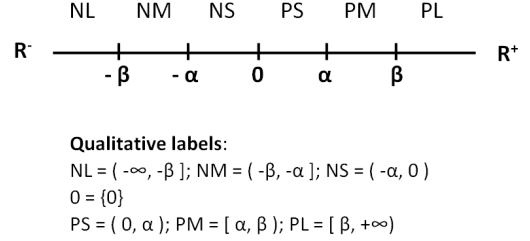

Qualitative labels:
NL = (-∞, -β]; NM = (-β, -α]; NS = (-α, 0)
0 = {0}
PS = (0, α); PM = [α, β); PL = [β, +∞)

Figure 1: The $OM(3)$ absolute partition.

Order of Magnitude Space

An order of magnitude (OM) space S defined for a qualitative variable is the combination of the ordered label set S_l and the interval-like treatment of such labels. For instance, the value of one variable is expressed by the set of basic labels $S_l = \{B_1, \ldots, B_n\}$ with $B_1 < \ldots < B_n$ denoting its qualitative order, meaning that $\alpha < \beta, \forall \alpha \in B_i, \beta \in B_j, i < j$. The corresponding OM space S is formally described as $S = S_l \cup \{[B_i, B_j] | B_i, B_j \in S_l, i < j\}$. In effect, the label $[B_i, B_j]$ with $i < j$ is defined as the union of the elements within the set $\{B_i, B_{i+1}, \ldots, B_j\}$. In addition, the order in S_l induces the partial order $\leq_p$ in S, which represents *being more precise than* or *being less general than*:

$$[B_i, B_j] \leq_p [B_p, B_q] \iff [B_i, B_j] \subset [B_p, B_q] \quad (1)$$

where $[B_i, B_i] = \{B_i\}$. According to Figure 2, the least precise label is $[B_1, B_n]$, denoted by ?. This manipulation of ordered labels allows reasoning and analysis with single or combined labels that may reflect uncertainty of one agent on another agent's judgement.
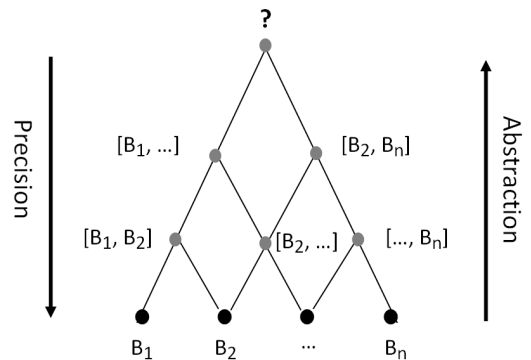

Figure 2: The graphical illustration of the partial order relation $\leq_p$ in an order-of-magnitude space S.

It is possible to define qualitative equality, termed q-equal, in an $OM(n)$ space S. Given $O, P \in S$, O and P are q-equal or $O \approx P$, if there is a $Q \in S$ such that $Q \leq_p O$ and $Q \leq_p P$. This effectively implies that O and P encompass, in part or in full, common basic elements. In addition, for presentational simplicity, $\forall O \in S$, the sets $B_O = \{B \in S_l - \{0\}, B \leq_p O\}$ and $B_O^* = \{B \in S_l, B \leq_p O\}$ are termed the *base of O* and the *enlarged base of O*, respectively.

Qualitative Algebra of AOM

At the outset, the mathematical structure of the AOM model, called Qualitative Algebra or Q-algebra, was initially defined as the unification of sign and interval algebra over a continuum of qualitative partitions of the real line (Travé-Massuyès & Piera 1989). However, although being superior to the sign algebra, such qualitative operators usually produce ambiguous and indeterminate outcomes. Accordingly, this barrier has been tackled via the notion of *qualitative expression of a real operator* (Agell, Rovira, & Ansotegui 2000). In particular, qualitative operators are considered as multidimensional functions defined in an AOM space. The Cartesian product of $S^1, S^2, \ldots, S^k$ (where k is the number of variables of a given problem domain, S^i is an $OM(n)$ space, $i = 1 \ldots k$) is adopted to express the outcome of a real operator in $\mathcal{R}^k$ qualitatively, which is reflected onto the resulting qualitative space S'.

Given a real operator ω defined on $\mathcal{R}^k$ involving k real variables with each taking values in $\mathcal{R}$, the corresponding qualitative abstraction of ω, denoted as $[\omega]$, is specified on S^k with values in S' as follows:

$$[\omega](X_1, X_2, \ldots, X_k) = [\omega(X_1, X_2, \ldots, X_k)]_{S'} \quad (2)$$

where $X_i \in S^i, i = 1 \ldots k$ and $\omega(X_1, X_2, \ldots, X_k) = \{\omega(x_1, x_2, \ldots, x_k), x_i \in X_i\}$. Inherently, $[\omega]$ assigns to each k-tuple element of $(X_1, X_2, \ldots, X_k)$ a qualitative description of the subset enclosing all underlying numerical results of applying ω over all real values in $X_1, X_2, \ldots, X_k$.

To simplify this, it is feasible to generate the qualitative operator, $[\omega]$, from the basic ordered labels of an OM space, $S, S^i = S, \forall i = 1 \ldots k$. For any $[\omega]$ and $X_1, X_2, \ldots, X_k \in S$:

$$[\omega](X_1, X_2, \ldots, X_k) = \bigcup_{B_i \in B_{X_i}^*} [\omega](B_1, B_2, \ldots, B_k) \quad (3)$$

According to Equation 2, the qualitative operator $[\omega]$ can be generalized as follows:

$$[\omega](X_1, \ldots, X_k) = \bigcup_{B_i \in B_{X_i}^*} [\omega(B_1, \ldots, B_k)]_{S'} \quad (4)$$

It is noteworthy that the $[\omega]$ operator presented above is compatible only to variables specified in the same order of magnitude space. To enhance the applicability of this terminology, the utilization of this qualitative operator is further introduced to multi-granularity domains via the homogenization of references, which has been successfully applied to realistic problems like credit risk prediction (Agell, Rovira, & Ansotegui 2000) and marketing segmentation (Olmo *et al.* 2007). This intuitive technique is extensively used in the current research, which will be thoroughly discussed below.

Order-of-Magnitude Based Link Analysis

This section introduces a novel order-of-magnitude based link analysis in which multiple link properties are combined to improve the quality of estimated link-based similarity measures.

Link Properties

Link analysis is based on examining relation patterns amongst references of real-world entities, which can be formally specified as an undirected graph $G(V, E)$. It is composed of two sets, the set of vertices V and that of edges E, respectively. Let X and R be the sets of all references and their relations in the dataset. Then, vertex $v_i \in V$ denotes reference $x_i \in X$ and each edge $e_{ij} \in E$ linking vertices $v_i \in V$ and $v_j \in V$ corresponds to a relation $r_{ij} \in R$ between references $x_i \in X$ and $x_j \in X$. Each edge $e_{ij} \in E$ possess statistical information $f_{ij} \in \{1, \ldots, \infty\}$, representing the frequency of any relation occurring between references x_i and x_j within the underlying dataset. With this terminology, several methods have been introduced to evaluate the similarity between information objects: SimRank (Jeh & Widom 2002), Connected-Triple (Klink *et al.* 2006), PageSim (Lin, King, & Lyu 2006) and a variety of random walk methods (Minkov, Cohen, & Ng 2006) (see more details in (Getoor & Diehl 2005) and (Liben-Nowell & Kleinberg 2007)).

Cardinality Property (CT) In essence, existing techniques, such as SimRank and Connected-Triple, have concentrated exclusively on the numerical count of shared neighboring objects. Let $v_i \in V$ be an entity of interest (e.g. a terrorist name in intelligence data or a paper in a publication database) and $N_{v_i} \subset V$ be a set of entities directly linked to v_i, called neighbors of v_i. The similarity between entities v_i and v_j is then determined by the cardinality of $N_{v_i} \cap N_{v_j}$, the set of shared neighbors where N_{v_i} and N_{v_j} are sets of neighbors of entities v_i and v_j, respectively. Effectively, the higher the cardinality is, the greater the similarity of these entities becomes.

Uniqueness Property (UQ) Despite their simplicity, cardinality based methods are greatly sensitive to noise and often generate a large proportion of false positives (Klink *et al.* 2006). This shortcoming emerges because these methods exclusively concern with the cardinality property of link patterns without taking into account the underlying characteristics of a link itself. As the first attempt to extend this approach by addressing such characteristics, the *uniqueness measure* of link patterns has been suggested as the additional

criterion to CT to refine the estimation of similarity values (Boongoen & Shen 2008).

Given a graph $G(V, E)$ in which objects and their relations are represented with members of the sets of vertices V and edges E, respectively, a uniqueness measure UQ_{ij}^k of any two objects i and j (denoted by vertices $v_i, v_j \in V$) can be approximated from each joint neighbor k (denoted by the vertex $v_k \in V$) as follows:

$$UQ_{ij}^k = \frac{f_{ik} + f_{jk}}{\sum_m f_{mk}} \quad (5)$$

where f_{ik} is the frequency of the link between objects i and k occurring in data, f_{jk} is the frequency of the link between objects j and k, and f_{mk} is the frequency of the link between object k and any object m.

To summarize the uniqueness of joint link patterns UQ_{ij} between objects i and j, the ratios estimated for each shared neighbor are aggregated as

$$UQ_{ij} = \frac{1}{n} \sum_{k=1}^n UQ_{ij}^k \quad (6)$$

where n is the number of overlapping neighbor objects that objects i and j are commonly linked to.

Link Based Similarity Evaluation

A common drawback of those numerical measures previously presented is the inability to achieve coherent and natural interpretation through existing seemingly fine-grained scales. Exploring a link network with crisp numerically-valued criteria is typically considered inflexible comparing to the use of interval and linguistic descriptors. Specifically, a wrong interpretation of a property measure may occur if there exists a unduly high property value within a link network. A more accurate and naturally expressive measure is to exploit qualitative labels like highly, moderately or poorly certain.

In order to overcome this important shortcoming, measures of link properties like cardinality and uniqueness are gauged in accordance with property-specific order-of-magnitude (OM) spaces. Subsequently, the link-based similarity value is derived by combining these qualitative descriptors each assigned with a possibly different degree of relevance. Homogenizing of references in multi-granularity OM spaces (Agell, Rovira, & Ansotegui 2000) is applied to this aggregation process in such a way that values measured in distinct scales can be analogously manipulated.

OM Spaces for Link Properties At the outset, measures of link properties, originally in quantitative terms, are translated into elements of ordered label sets. Formally, let P^i and L^i be the set of intervals partitioned on the real line and that of the corresponding qualitative labels, defined for measures of the link property i on the discourse U^i. That is, $P^i = \{p_1^i, \ldots, p_{n^i}^i\}$ and $L^i = \{l_1^i \ldots l_{n^i}^i\}$, where n^i is the number of intervals/labels and $l_1^i < \ldots < l_{n^i}^i$ denotes the qualitative orders of magnitude specified for property i. Without causing confusion, for simplicity, intervals partitioned on real number line are termed partitions. They

are non-overlapped over the discourse U^i, and their crisp boundaries are determined by one or two members of the landmark set $M^i = \{m_1^i, \ldots, m_{n^i-1}^i\}$. Each partition p_j^i is qualitatively expressed by the label $l_j^i, \forall j = 1 \ldots n^i$, and its interval is defined by lower bound α_j^i and/or upper bound β_j^i such that $\alpha_j^i, \beta_j^i \in M^i$ and $\alpha_j^i \leq \beta_j^i$.

Intuitively, the number of labels should be small enough so as not to impose useless precision onto analysts, but it must be rich enough to allow meaningful assessment and discrimination of measurement (Herrera & Herrera-Viedma 2000). In fact, average human beings can reasonably manage to bear in mind seven or so items/labels (Miller 1956).

For the current research with $i \in \{CT, UQ\}$, as a simple example, measures of the cardinality property over the discourse $U^{CT} = [0, \infty)$ may be described using a member of the label set of three qualitative labels ($n^{CT} = 3$), $L^{CT} = \{l_1^{CT} = Small, l_2^{CT} = Medium, l_3^{CT} = Large\}$. In particular, if the landmark set $M^{CT} = \{m_1^{CT} = 2, m_2^{CT} = 6\}$, members of the partition set are specified as $P^{CT} = \{p_1^{CT} = [0, 2], p_2^{CT} = (2, 6], p_3^{CT} = (6, \infty)\}$. Likewise, the uniqueness measure, whose values can be defined on the universe of discourse $U^{UQ} = [0, 1]$, which may be expressed using the ordered set of five qualitative descriptors ($n^{UQ} = 5$), $L^{UQ} = \{l_1^{UQ} = VeryLow, l_2^{UQ} = Low, l_3^{UQ} = Moderate, l_4^{UQ} = High, l_5^{UQ} = VeryHigh\}$. Using the set of landmarks ($M^{UQ} = \{m_1^{UQ} = 0.1, m_2^{UQ} = 0.3, m_3^{UQ} = 0.6, m_4^{UQ} = 0.8\}$), the corresponding partition set can be defined as $P^{UQ} = \{p_1^{UQ} = [0, 0.1], p_2^{UQ} = (0.1, 0.3], p_3^{UQ} = (0.3, 0.6], p_4^{UQ} = (0.6, 0.8], p_5^{UQ} = (0.8, 1]\}$.

Similarity Measure via Aggregation of Properties Relying on one particular link property, as with existing link-based methods, for justifying the similarity between any two objects in a link network may lead to false interpretation and perhaps revolting consequences. The more rational alternative is to integrate all available link properties in order to refine the similarity measure. Fortunately, the link-based similarity between any two vertices $v_a, v_b \in V$ in the link network can be estimated through the aggregation of qualitative descriptors each corresponding to a particular link property i. In particular, each property i can be assigned with a different degree of relevance (e.g. importance) RV^i, which may be given by domain experts in according with their past experiences or estimated from past data if such expertise is not readily available. Similar to measures of link properties previously emphasized, relevance can be naturally expressed using the order-of-magnitude label set L^{RV}, such as $L^{RV} = \{None, +, ++, +++\}$ or $L^{RV} = \{0, 1, 2, 3\}$. In the discussion above, the relevance degrees of cardinality $RV^{CT} \in L^{RV}$ and uniqueness properties $RV^{UQ} \in L^{RV}$ are subjectively set to 2 and 1, respectively.

However, since label sets defined for different properties are usually of unequal granularity, they have to be homogenized onto a common scale on which references of distinct label sets can be uniformly manipulated and integrated. Following the work of (Agell, Rovira, & Ansotegui 2000), the

Table 1: Homogenized landmarks.

Landmarks	CT	UQ
Original	2, 6	0.1, 0.3, 0.6, 0.8
Step1	0, 4	-0.2, 0, 0.3, 0.5
Step2	-4, 0, 4	-0.5, -0.3, -0.2, 0, 0.2, 0.3, 0.5
Step3	-4, -2, -1, 0, 1, 2, 4	-0.5, -0.3, -0.2, 0, 0.2, 0.3, 0.5
Homogenized	-3, -2, -1, 0, 1, 2, 3	-3, -2, -1, 0, 1, 2, 3
Irrelevant	-3, -2, -1, 1, 2	-3, -2, 1

procedure below will be used here:

- *Step1*: Convert each set of landmarks M^i into a symmetric arrangement. Given a central landmark $m_c^i \in M^i$, translate each landmark $m_t^i, t = 1 \ldots n^i - 1$ to the new landmark sm_t^i in the symmetric scale using $sm_t^i = m_t^i - m_c^i$. Note that the central landmark is now 0 in the new scale.

- *Step2*: Landmarks appearing on both positive and negative sides may be dissimilar in general. A fully symmetric pattern can be achieved by adding missing landmarks, so that one absolute landmark can be found on both positive and negative sides of 0. Obviously, these newly added elements are of balancing purpose only, therefore they will not be used to represent values and will be deliberately marked as irrelevant.

- *Step3*: The landmark sets for each property are further modified by adding new landmarks on both side of 0, in such a way that all landmark sets have the same cardinality. Similar to Step 2, new elements are irrelevant with respect to each particular property and are simply to support the unification mechanism.

In accordance to the landmarks of two link properties given earlier, Table 1 summarizes the results achieved at each step of the homogenization process.

Following the terminology of AOM algebra, with the property-specific relevance degrees previously clarified, order-of-magnitude based similarity measure (OMS) can be estimated from measures of any n properties using the qualitative expression of a real weighted summation $[\omega]$:

$$OMS = [\omega](X_1, \ldots, X_n, RV_1, \ldots, RV_n)$$
$$= [\omega(X_1, \ldots, X_n, RV_1, \ldots, RV_n)]_{S^{Sum}} \quad (7)$$

where $X_i \in S^H$ is the qualitative measure of link property $i, i = 1 \ldots n$, expressed on the homogenized scale S^H, RV_i is its corresponding relevance degree, S^{Sum} is the resulting order-of-magnitude space of this summarization and ω is defined as

$$\omega(X_1, \ldots, X_n, RV_1, \ldots, RV_n) = \omega(x_1, \ldots, x_n, rv_1, \ldots, rv_n)$$
$$= x_1 rv_1 + \ldots + x_n rv_n \quad (8)$$

where $x_i \in X_i, rv_i \in RV_i, i = 1 \ldots n$.

Specific to the two link property measures used herein: CT and UQ, with their relevance degrees being RV^{CT} and RV^{UQ} and the homogenized scale S^H being {-3, -2, -1, 0, 1, 2, 3}, the previous equations can be employed as follows:

$$OMS = [\omega](CT, UQ, RV^{CT}, RV^{UQ})$$
$$= [\omega(CT, UQ, RV^{CT}, RV^{UQ})]_{S^{Sum}} \quad (9)$$

Following that

$$OMS = [\omega(2ct + uq)]_{S^{Sum}} \quad (10)$$

where $ct \in M^{CT}$, and M^{CT} is the set of relevant landmarks of CT in the homogenized scale S^H: $M^{CT} = \{0, 3\}$. Likewise, uq is a member of M^{UQ}, with $M^{UQ} = \{-1, 0, 2, 3\}$. Effectively, the resulting order-of-magnitude space S^{Sum} is established upon landmark values of this qualitative operation, which are $\{-1, 0, 2, 3, 5, 6, 8, 9\}$. To obtain a coherent interpretation of similarity measures within the S^{Sum} space, a set of qualitative labels L^{OMS}, as partitions of S^{Sum}, is chosen to express the different orders of magnitude of the similarity values. For instance, $L^{OMS} = \{Low\ (OMS < 2), Medium\ (2 \le OMS \le 6), High\ (OMS > 6)\}$. Note that a more or less refined label sets can be used depending on the precision level required.

Semi-Supervised Method to Designing Landmarks

Designing an appropriate set of landmarks M^i for a link property i is non-trivial and proves to be critical towards the quality of generated similarity measures. A simple approach is to rely on human experts, who select suitable landmark values in accordance with their personal intuition and judgment. This is not usually effective regarding the availability of experts and the diverse nature of different problem domains. Besides, human input may be rather subjective and inconsistent. As a result, a data-driven mechanism that can be used to obtain an appropriate M^i is specifically discussed herein.

For a link property i, a density graph is formulated to represent the proportion of entity pairs (i.e. (v_x, v_y), $v_x, v_y \in V$), each with different property measure i_{xy}. Let $D : [0, i_{max}] \to [0, 1]$ be the density function (where i_{max} denotes the maximum value of i_{xy}), which is formally defined as

$$D(t) = \frac{N(t)}{\sum\limits_{\forall r \in [0, i_{max}]} N(r)} \qquad (11)$$

where $N(t)$ denotes a number of entity pairs (v_x, v_y) whose property measure $i_{xy} \geq t, t \in [0, i_{max}]$. Figure 3 presents the density function of cardinality property (i.e. $i = CT$) derived from the Terrorist dataset (Hsiung *et al.* 2005), where $CT_{max} = 113$ (and the magnified presentation of $D(t), t \in \{7, 113\}$ is included for better interpretation).
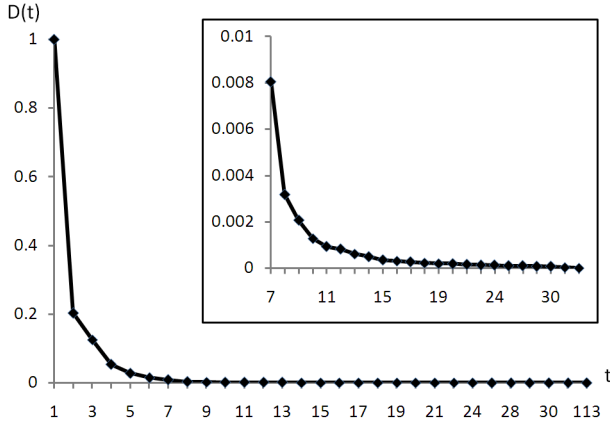

Figure 3: Example of density function derived from Terrorist dataset.

With this function, the following set of heuristics can be articulated especially to help data analysts to assess a proper set of landmarks M^i for link property i:

- Let $M^i = \{m_1^i, m_2^i, \ldots, m_{n^i}^i\}$ be an appropriate landmark set for property i, where $m_g^i \leq i_{max}, \forall g \in \{1 \ldots n^i\}$ and $m_h^i \leq m_{h+1}^i, \forall h \in \{1 \ldots n^i - 1\}$.

- Each pair of adjacent landmarks (i.e. m_h^i and m_{h+1}^i) encapsulates all property values $i_{xy} \in [m_h^i, m_{h+1}^i)$ whose density $D(i_{xy})$ can be perceived at a particular order of magnitude. Note that orders of magnitude utilized in this research are of $\alpha \times 10^z$, where $z \in \{-1, -2, \ldots, -\infty\}$ and $\alpha \in (0, 10)$. According to Figure 3, M^{CT} of the Terrorist dataset is $\{4, 7, 10, 23\}$ such that $D(CT_{xy})$ is expressed at five different orders of magnitude of

  - 10^{-1} where $CT_{xy} < 4$
  - 10^{-2} where $4 \leq CT_{xy} < 7$
  - 10^{-3} where $7 \leq CT_{xy} < 10$
  - 10^{-4} where $10 \leq CT_{xy} < 23$
  - 10^{-5} where $CT_{xy} \geq 23$

This semi-supervised method is effective to assist analysts to design appropriate landmarks and descriptive labels, based on quality measures of the particular link network being studied. Unlike human-directed alternatives, it is data oriented and capable of being adapted to a variety of problems.

Application to False Identity Detection

This section presents the application of the order-of-magnitude link-based similarity evaluation to detecting the use of false identities. Particularly, its performance is empirically evaluated over the terrorism-related dataset, and further generalized with a publication data collection.

False Identity Detection

To battle false identity, an exact-match query to a law enforcement computer system is simply ineffective. A better approach extensively studied in (Bilenko & Mooney 2003) and (Wang *et al.* 2006) is to exploit the similarity measure of names obtained from one or several string-matching techniques. Despite their reported success, these *content-based* methods can not handle cases where completely different names are deployed. For instance, they would fail to recognize the association between these pairs of terrorists' name, whose overlapping text content is void.

- (*ashraf refaat nabith henin, salem ali*)
- (*bin laden, the prince*)
- (*bin laden, the emir*)
- (*abu mohammed nur al-deen, the doctor*)

Accordingly, the *link-based* approach, which has proven effective for similar problems in a wide range of domains (e.g. publication (Klink *et al.* 2006), online resources (Hou & Zhang 2003), (Lin, King, & Lyu 2006), email (Minkov, Cohen, & Ng 2006) and intelligence data analysis (Hsiung *et al.* 2005)), has been put forward to underpin the accountability for unstructured information.

Let O be the set of real-world entities each being referred to by at least one member of another set X, which is a collection of names or references. A pair of names (x_i, x_j) are aliases when both names correspond to the same real-world entity: $(x_i \equiv o_k) \wedge (x_j \equiv o_k), o_k \in O$. In practice, disclosing an alias pair in graph G is to find a couple of vertices (v_i, v_j), whose similarity $s(v_i, v_j)$ is significantly high. Intuitively, the higher $s(v_i, v_j)$ the greater the possibility that vertices v_i and v_j, and hence corresponding names x_i and x_j, constitute the actual alias pair.

Datasets

The performance and applicability of the proposed approach is evaluated over the following distinct datasets: Terrorist (Hsiung *et al.* 2005) and DBLP (Klink *et al.* 2006). Terrorist is a link dataset manually extracted from web pages and news stories related to terrorism. Each node presented in this link network is a name of person, place or organization, while a link denotes a co-occurrence association between objects through reported events. Figure 4 presents an example of this link network where names *Bin laden* and *Abu abdallah* refer to the same real-world person.

DBLP (Digital Bibliography and Library Project) is the dataset containing co-authoring information extracted from different bibliographical databases. In this link network, each node represents a reference name of an author and a link denotes the fact that two names appear as the co-authors
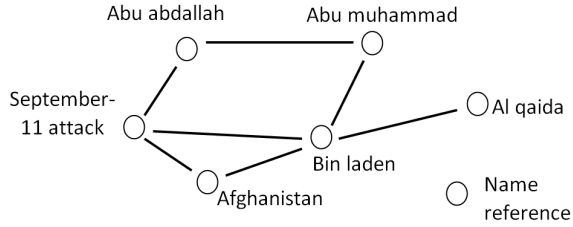
Figure 4: An example of Terrorist dataset.

of a paper (or papers). Table 2 summarizes the number of links, objects and alias pairs included in these datasets.

Table 2: Dataset details (number of objects, links and alias pairs).

Dataset	Objects	Links	Alias Pairs
Terrorist	4088	5581	919
DBLP	2796	8157	23

Performance Evaluation

Efficiency of Semi-Supervised Method Initially, it is important to examine the effectiveness of the proposed semi-supervised method for modeling a landmark set. By following the heuristics previously prescribed, appropriate landmark values are:

- For Terrorist dataset, $M^{CT} = \{4, 7, 10, 23\}$ and $M^{UQ} = \{0.05, 0.12, 0.27, 0.43, 1\}$.

- For DBLP dataset, $M^{CT} = \{2, 5, 9, 15\}$ and $M^{UQ} = \{0.008, 0.04, 0.17, 0.31, 1\}$.

With these data-oriented landmarks, Table 3 compares the number of disclosed alias pairs successfully detected by different methods, where K denotes the number of entity pairs with highest similarity measures (details of homogenization for semi-supervised landmarks are not included due to space limitation). Note that OMS and OMS^H represent order-of-magnitude based similarity measures, with semi-supervised and human-directed landmarks, respectively. In addition, QT denotes a simple integration of numerical CT and UQ measures, where relevance degrees RV^{CT} and RV^{UQ} (2 and 1, respectively) similar to those of OMS and OMS^H are employed.

These results indicate that the OMS measure with semi-supervised landmarks usually outperforms both human-directed landmarks OMS^H and the quantitative evaluation QT, especially over Terrorist dataset.

Comparison with Alternative Link-Based Methods The performance of the OMS method is further generalized by evaluating against the following two state-of-the-art link-based measures: SimRank (SR) and PageSim (PS), respectively.

Table 3: Number of alias pairs disclosed by each method.

K	OMS	OMS^H	QT
Terrorist			
200	43	9	8
400	80	57	41
600	115	91	60
800	146	110	75
1000	180	138	102
DBLP			
100	4	1	1
200	5	2	1
300	5	3	2
400	6	5	4
500	10	6	5

- Principally, with the objective of finding similar publications given their citation relations, SimRank relies on the cardinality of shared neighbors that are iteratively refined to a fixed point (Jeh & Widom 2002). In each iteration, the similarity of any pair of vertices $v_i, v_j \in V$, $s(v_i, v_j)$, is approximated as

$$s(v_i, v_j) = \frac{C \sum_{p=1}^{|N_{v_i}|} \sum_{q=1}^{|N_{v_j}|} s(N_{v_i}^p, N_{v_j}^q)}{|N_{v_i}||N_{v_j}|} \quad (12)$$

where $N_{v_i}, N_{v_j} \subset V$ are sets of neighboring vertices to which vertices v_i and v_j are linked, respectively. Individual neighbors of both vertices are denoted as $N_{v_i}^p$ and $N_{v_j}^q$, for $1 \leq p \leq |N_{v_i}|$ and $1 \leq q \leq |N_{v_j}|$. The constant $C \in [0, 1]$ is a decay factor that represents the confidence level of accepting two non-identical entities to be similar. Note that $s(v_i, v_j) = 0$ when $N_{v_i} = \emptyset$ or $N_{v_j} = \emptyset$.

- Within a different domain, PageSim (Lin, King, & Lyu 2006) was developed to capture similar web pages based on associations implied by their hyperlinks. In essence, the similarity measure $ps(v_i, v_j)$ between vertices v_i and v_j is dictated by the coherence of ranking scores $R(v_g, v_i)$ and $R(v_g, v_j)$ propagated to them from any other vertex $v_g \in V$. It is noteworthy that ranking scores are explicitly generated using the page ranking scheme, PageRank (Brin & Page 1998), of the most developed Google search engine (with detailed computational mechanism for the ranking scores omitted here). Formally, PageSim can be defined as

$$ps(v_i, v_j) = \sum_{\forall v_g \in V, v_g \notin \{v_i, v_j\}} \frac{\min(R(v_g, v_i), R(v_g, v_j))^2}{\max(R(v_g, v_i), R(v_g, v_j))} \quad (13)$$

According to Table 4, the OMS measure consistently outperforms other link-based methods over both datasets. In spite of its low performance, the SimRank measure, which

has been recognized as a benchmark link analysis technique for publication (Getoor & Diehl 2005) and Internet (Calado *et al.* 2006) domains, is included in this evaluation as to reflect the difficulty of this task. Based on the results presented in Tables 3-4, the proposed method does encounter the problem of false positives. However, its performance with respect to this difficulty has been substantially improved as compared to other link-based similarity methods.

Table 4: Number of alias pairs disclosed by each method.

K	OMS	SR	PS
Terrorist			
200	43	0	7
400	80	0	36
600	115	1	63
800	146	1	79
1000	180	2	92
DBLP			
100	4	0	1
200	5	1	1
300	5	2	1
400	6	2	2
500	10	3	4

Computational Complexity In addition to evaluating these methods in terms of discovered alias pairs, it is important to investigate the computational complexity that would determine or even limit their actual real-world applications. Let a link network consist of n distinct entities, each averagely linked to other m entities. The time complexity for the OMS approach to generate all pair-wise similarity values is $O(n^2m^2)$. With f iterations of similarity refinement, the time complexity of SimRank is $O(n^2m^2f)$. Note that the results shown in Table 4 are obtained using $f = 3$ (with its usual range being 3-5).

In contrast, the PageSim is rather complex compared to the others as it begins with ranking all entities using the PageRank technique, whose time complexity is $O(nmt)$ where t is the number of iterations for refining the ranking values (with t being 3 in this experiment). Having accomplished the ranking process, the similarity of two entities is estimated on the ranking values propagated from their shared neighbors, with the maximum connecting-path length of r (r set to 3 for the results given in Table 4). As a result, the overall time complexity of PageSim method is $O(n^2m^{2r} + nmt)$.

Hence, the OMS method introduced in this paper not only performs well in terms of precision, but also proves to be practical for alias detection, with efficient time consumption.

Conclusion

This paper has presented a novel qualitative link-based similarity measure, which can be efficiently employed for intelligence data analysis and disclosing the use of false identity typically appearing in terrorists and criminals' activities. Unlike initial numerical similarity estimation that concentrates solely on the link structures, the qualitative method also includes underlying link properties such as uniqueness in order to purify the similarity description. In addition, qualitatively distinct order-of-magnitude labels incorporate semantics towards similarity justification and allow coherent interpretation and reasoning that is hardly feasible with pure numerical terms.

Technically, measures of link properties are gauged in accordance with property-specific order of magnitude spaces, whose dissimilar scales are subsequently homogenized to permit the unification of their values. In essence, the similarity descriptor is achieved via aggregating property values regarding to their relative degrees of relevance. Empirically, this qualitative approach consistently outperforms numerical similarity measures over terrorism-related and publication datasets. However, in order to generalize its performance and applicability, it is crucial to evaluate this method with more relevant data. Also, relevance degrees allocated for distinct link properties may be better learned from data, instead of relying on human-directed ones.

Acknowledgments

References

Agell, N.; Rovira, X.; and Ansotegui, C. 2000. Homogenising references in orders of magnitude spaces: An application to credit risk prediction. In *Proceedings of International Workshop on Qualitative Reasoning*, 1–8.

Badia, A., and Kantardzic, M. M. 2005. Link analysis tools for intelligence and counterterrorism. In *Proceedings of IEEE International Conference on Intelligence and Security Informatics, Atlanta*, 49–59.

Bilenko, M., and Mooney, R. J. 2003. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 39–48.

Boongoen, T., and Shen, Q. 2008. Detecting false identity through behavioural patterns. In *Proceedings of Int. Crime Science Conference, London*.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30(1-7):107–117.

Calado, P.; Cristo, M.; Gonçalves, M. A.; de Moura, E. S.; Ribeiro-Neto, B. A.; and Ziviani, N. 2006. Link based similarity measures for the classification of web documents. *Journal of American Society for Information Science and Technology* 57(2):208–221.

Getoor, L., and Diehl, C. P. 2005. Link mining: a survey. *ACM SIGKDD Explorations Newsletter* 7(2):3–12.

Herrera, F., and Herrera-Viedma, E. 2000. Linguistic decision analysis: steps for solving decision problems under linguistic information. *Fuzzy Sets and Systems* 115:67–82.

Hou, J., and Zhang, Y. 2003. Effectively finding relevant web pages from linkage information. *IEEE Transactions on Knowledge and Data Engineering* 15(4):940–951.

Hsiung, P.; Moore, A.; Neill, D.; and Schneider, J. 2005. Alias detection in link data sets. In *Proceedings of International Conference on Intelligence Analysis*.

Jeh, G., and Widom, J. 2002. Simrank: A measure of structural-context similarity. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 538–543.

Klink, S.; Reuther, P.; Weber, A.; Walter, B.; and Ley, M. 2006. Analysing social networks within bibliographical data. In *Proceedings of International Conference on Database and Expert Systems Applications*, 234–243.

Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58(7):1019–1031.

Lin, Z.; King, I.; and Lyu, M. R. 2006. Pagesim: A novel link-based similarity measure for the world wide web. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, 687–693.

Miller, G. 1956. The magical number seven, plus or minus two: Some limits on our capacity for information processing. *Psychological Review* 63(2):81–97.

Minkov, E.; Cohen, W. W.; and Ng, A. Y. 2006. Contextual search and name disambiguation in email using graphs. In *Proceedings of Int. Conference on Research and Development in IR*, 27–34.

Olmo, C.; Sánchez, G.; Prats, F.; Agell, N.; and Sánchez, M. 2007. Using orders of magnitude and nominal variables to construct fuzzy partitions. In *Proceedings of IEEE International Conference on Fuzzy Systems*, 1–6.

Piera, N. 1995. Current trends in qualitative reasoning and applications. *Monografia CIMNE, 33. International Center for Numerical Methods in Engineering, Barcelona*.

Porter, G. 2008. Crying (iranian) wolf in argentina. *Asia Times Online (www.atimes.com)*.

Travé-Massuyès, L., and Piera, N. 1989. The orders of magnitude models as qualitative algebras. In *Proceedings of 11th International Joint Conference on Artificial Intelligence*, 1261–1266.

Wang, A. G.; Atabakhsh, H.; Petersen, T.; and Chen, H. 2005. Discovering identity problems: A case study. In *Proceedings of IEEE International Conference on Intelligence and Security Informatics, Atlanta*, 368–373.

Wang, G. A.; Chen, H.; Xu, J. J.; and Atabakhsh, H. 2006. Automatically detecting criminal identity deception: an adaptive detection algorithm. *IEEE Transactions on Systems, Man and Cybernetics, Part A* 36(5):988–999.