Semi-Supervised OWA Aggregation for Link-Based Similarity Evaluation and Alias Detection

Shen, Qiang; Boongoen, Tossapon

DOI:
10.1109/FUZZY.2009.5277168

Publication date:
2009

Citation for published version (APA):
Shen, Q., & Boongoen, T. (2009). *Semi-Supervised OWA Aggregation for Link-Based Similarity Evaluation and Alias Detection*. 288-293. https://doi.org/10.1109/FUZZY.2009.5277168

Semi-Supervised OWA Aggregation for Link-Based Similarity Evaluation and Alias Detection

Tossapon Boongoen and Qiang Shen

Abstract— **Within the past decades, many fuzzy aggregation techniques, ordered weighted averaging (OWA) in particular, have proven effective for a wide range of information processing tasks, such as decision making, image analysis, database and machine learning. Despite reported successes, their potentials have yet to be explored for the emerging problem of link analysis, which aims to discover similarity and relations amongst objects through their associations. Recently, several link-based similarity methods have been put forward to identifying similar objects in the Internet and publication domains. However, these techniques only take into account the cardinality property of a link structure that is highly sensitive to noise and causes a great number of false positives. In light of such challenge, this paper presents a novel OWA aggregation model that is capable of efficiently deriving a similarity measure through the integration of multiple link properties. The underlying approach is based on the methodology of stress function by which the aggregation behavior can be easily interpreted and modeled. In addition, a semi-supervised method is introduced to assist a user in designing a stress function, i.e. the weighting scheme of link properties, appropriate for a particular link network. The application of the OWA aggregation approach to alias detection is demonstrated and evaluated, against state-of-art link-based techniques, over datasets specifically related to terrorism, publication and email domains.**

I. Introduction

Within the past decades, many aggregation techniques have been invented for a variety of fuzzy information processing tasks [1], [18]. Particularly to the ordered weighted averaging (OWA) operator [21], it has been applied to a vast number of fields since its introduction in 1988, including fuzzy logic controller [22], market analysis [25], image compression [16], query system [20], feature selection [2] and decision making [4], [21]. In spite of this, its potential has yet to be investigated for the emerging problem of link analysis (i.e. link mining) [9], [13].

As a response to the increasing amount of link oriented information (e.g. online resources), many link analysis techniques have been developed to disclose similarity amongst objects through the pattern of their relations [13]. They are effective to overcome the fundamental pitfalls of conventional text-based methods, which require large storage and long computing time due to the need of full-text comparison [6], [14]. The advantages of link analysis have been especially recognized for problems such as the World Wide Web where text-based methods are sometimes inapplicable with pages containing little texts but a large amount of multimedia objects [14], and for intelligence data analysis where content-based approaches can be misleading due to fraud descriptions of terrorists' name, appearance and contact details [3], [19].

Essentially, several link-based methods, such as SimRank [11], Connected-Triple [12] and PageSim [14], analogously justify the similarity between any two objects in a link network upon the cardinality of joint neighbors to which these two are linked. Despite its simplicity, by not taking into account other link characteristics, this measure is sensitive to noise and usually causes a great number of false positives. However, the quality of the similarity evaluation may be enhanced by including uniqueness aspect of links within the overlapping neighbor context [3].

Recognizing the aforementioned shortfall, this paper presents a new OWA aggregation model that is efficient to derive link-based similarity measure through the integration of multiple properties of a link pattern. With the methodology of stress function [24], the underlying aggregation behavior, and hence the weighting scheme of link properties, can be effectively modeled and comprehended. Yet, a human-directed stress function can not be uniformly formulated for a variety of problems, whose characteristics can vary greatly. As a result, a semi-supervised mechanism is introduced to assist analysts to obtain an appropriate function. Particularly, a set of heuristics with graphical projection of link measures are provided to support the relevant analysis process.

The rest of this paper is organized as follows. Section 2 introduces the OWA aggregation with stress function, upon which the present research is developed. Following that, Section 3 presents the motivation and details of the OWA aggregation model for link-based similarity evaluation, especially different link properties and stress functions used in this aggregation process. The forth section describes the semi-supervised method to assist data analysts to design an appropriate stress function. Section 5 details the experimental evaluation of the application of the present work to detecting alias names or duplicates in different datasets. The paper is concluded in Section 6, with the perspective of further work.

II. OWA Aggregation with Stress Function

The process of information aggregation appears in many application problems. Despite computationally simplistic, neither minimum nor maximum may be appropriate for many such applications. Accordingly, Yager [21] pioneered a new set of aggregation techniques called the ordered weighted averaging (OWA) operator. This mean-type operator provides a flexibility to utilize the entire range of *and* to *or* associated with the actual scenario in which information aggregation is

Tossapon Boongoen and Qiang Shen are with the Department of Computer Science, Aberystwyth University, UK (email: {tsb,qqs}@aber.ac.uk).

978-1-4244-3597-5/09/$25.00 ©2009 IEEE

required. As the current research serves as an initial investigation of bridging the OWA aggregation with link analysis approach, its fundamental concepts and weight determination methods are specifically emphasized herein.

An OWA operator of dimension n is a mapping $OWA : R^n \rightarrow R$, which has an associated weighting vector $W = (w_1, w_2, \ldots, w_n)^T$, where $w_i \in [0,1]$ and $\sum_{i=1}^{n} w_i = 1$. Given an input argument vector $X = (x_1, x_2, \ldots, x_n) \in R^n$, OWA is defined as follows:

$$OWA(X) = OWA(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} w_i x_{\sigma(i)} \quad (1)$$

where σ is a permutation function that orders the elements such that $x_{\sigma(i)} \geq x_{\sigma(i+1)}, \forall i = 1, \ldots, n-1$.

Weight determination is crucial to this family of operators, since associated weights dictate the type of aggregation operator an OWA exhibits. A number of different techniques have been proposed for obtaining weights associated with the OWA operator, for instance, maximal entropy [17], weight learning [7] and data clustering methods [2] (see more details in [8]). In addition, another important and useful approach to weight determination is through the manipulation of a certain type of function [23], which is outlined below.

Let F be a function $F : [0,1] \rightarrow [0,1]$ such that $F(0) = 0$, $F(1) = 1$ and $F(a) \geq F(b)$ given $a \geq b$. Using this function, it is possible to derive a weight vector $(w_1, w_2, \ldots, w_n)^T$ as

$$w_i = F\left(\frac{i}{n}\right) - F\left(\frac{i-1}{n}\right), i = 1 \ldots n \quad (2)$$

Following this, Yager [24] recently introduced a simple weight generation mechanism with stress functions, by which a user can conceptually specify the type of OWA operator required for a particular problem. Here, stress reflects significance. In particular, a *stress* function is a non-negative function $s(x)$ defined on the unit interval $x : [0,1] \rightarrow R^+$ (see Figure 1 for examples). Given this function, $F(x)$ can be defined as follows, where $\int_0^1 s(y)dy = K$:

$$F(x) = \frac{1}{K} \int_0^x s(y)dy \quad (3)$$

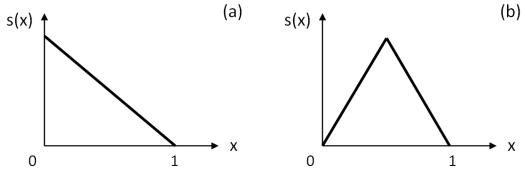

Fig. 1. Examples of stress function from [24]: (a) linearly decreasing and (b) centering-type linear functions.

According to Equation 2, OWA weights can be derived as

$$w_i = \frac{1}{K}\left(\int_0^{\frac{i}{n}} s(y)dy - \int_0^{\frac{i-1}{n}} s(y)dy\right), i = 1 \ldots n \quad (4)$$

In essence, this calculation can be simplified as weights being approximated directly from a stress function s as (see proof and further details in [24])

$$w_i = \frac{s\left(\frac{i}{n}\right)}{\sum_{i=1}^{n} s\left(\frac{i}{n}\right)} \quad (5)$$

With this method, consistent weight vectors can be obtained for different argument cardinalities and a user can easily characterize the nature of aggregation through locations of stress.

III. OWA Aggregation Model for Link-Based Similarity Evaluation

This section introduces a novel OWA aggregation framework in which multiple link properties are combined to improve the quality of estimated link-based similarity measures.

A. Link Properties for Link-Based Similarity Assessment

The link analysis approach 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 [11], Connected-Triple [12], PageSim [14] and a variety of random walk methods [15] (see more details in [9] and [13]).

1) 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.

2) Uniqueness Property (UQ): Despite their simplicity, cardinality based methods are greatly sensitive to noise and often generate a large proportion of false positives [12]. 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 [3].

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:

$$UQ_{ij}^k = \frac{f_{ik} + f_{jk}}{\sum_m f_{mk}} \tag{6}$$

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

$$UQ_{ij} = \frac{1}{n} \sum_{k=1}^{n} UQ_{ij}^k \tag{7}$$

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

B. Integrating Multiple Link Properties using OWA Aggregation with Stress Function

Intuitively, the similarity of objects $v_i, v_j \in V$ in a link network $G(V, E)$ may be justified with the cardinality (CT) of their joint neighbors or the average uniqueness of links (UQ) to the common peers. However, neither of these proves to be effective for all, various domain-specific, real-world data. Hence, to achieve a robust and accurate similarity estimation model, these link measures are proficiently integrated such that their significance degrees (i.e. weights) can be determined for an enhanced performance.

For this purpose, the methodology of stress function [24] appears appropriate to perform the aggregation if a data analyst can conveniently dictate the aggregating behavior through a graphical presentation of stress. Using the stress functions given in Figure 2, the similarity $s(v_i, v_j) \in [0, 1]$ of objects $v_i, v_j \in V$ can be approximated as follows:

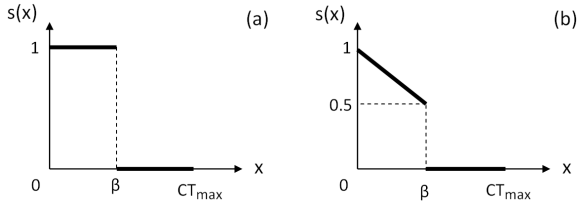

Fig. 2. Stress functions exploited in the similarity estimation process: (a) Stress1, constrained constant and (b) Stress2, constrained decreasing.

- *Step1*: Let CT_{ij} and $UQ_{ij} = \{UQ_{ij}^1 \dots UQ_{ij}^{CT_{ij}}\}$ be the number of common neighbors between v_i and v_j, and the set of uniqueness values measured at each of these shared objects (see Equation 6). Also let $|UQ_{ij}|$ denote the cardinality of the uniqueness set, which initially equals to CT_{ij}. Select the value of $\beta \in \{1, 2, \dots, CT_{max}\}$ (where CT_{max} is the maximum value of CT_{ij} in the studied dataset) that represents the number of member values in UQ_{ij} required for the efficient estimation of similarity, based on past experience or data-directed guideline. *Essentially, the value of β determines the weighting scheme exploited to combine CT and UQ measures.* The higher β is, the greater the significance of the CT measure becomes, as compared to the UQ counterpart. The parameterized stress function not only allows this aggregation model to be adapted to a wide range of problems, but also provides a practical means by which users can encode their attitude towards the aggregation of link measures.
- *Step2*: For each uniqueness set UQ_{ij}, if $|UQ_{ij}| < CT_{max}$, additional members with a zero uniqueness value are appended to UQ_{ij} such that $|UQ_{ij}|$ becomes CT_{max}. This process is to ensure that the similarities of all object pairs (v_i, v_j) with a different cardinality of UQ_{ij} are uniformly evaluated, through the same weighting scheme (i.e. β value).
- *Step3*: Estimate the similarity $s(v_i, v_j)$ by aggregating member values of the corresponding uniqueness set UQ_{ij} as

$$s(v_i, v_j) = OWA(UQ_{ij}^1, \dots, UQ_{ij}^{CT_{max}}) \tag{8}$$

For the current research, the stress functions shown in Figure 2 are exploited to generate weight vectors for the aforementioned OWA aggregation. In particular, given a stress function that is defined by $s : \{0, \dots, CT_{max}\} \rightarrow [0, 1]$, a weight vector $W = \{w_1, \dots, w_{CT_{max}}\}$ can be acquired as follows:

$$w_t = \frac{s(t)}{\sum_{r=1}^{CT_{max}} s(r)},\ t = 1 \dots CT_{max} \tag{9}$$

It is noteworthy that $w_t = 0, \forall t > \beta$.

IV. Semi-Supervised Method to Designing Stress Function for Link Analysis

Designing a stress function for the proposed OWA aggregation model, i.e. selecting a value of $\beta \in \{1, \dots, CT_{max}\}$, 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 pick up a suitable value, $\beta = CT_{max}$ for instance, 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 assist an analyst to obtain an appropriate β is specifically discussed herein.

At the outset, a density graph is formulated to represent the proportion of entity pairs (i.e. (v_i, v_j), $v_i, v_j \in V$) with different cardinality measure (CT). Let $D : \{1 \dots CT_{max}\} \rightarrow [0, 1]$ be the density function, which is formally defined as

$$D(t) = \frac{N(t)}{\sum_{r=1\ldots CT_{max}} N(r)},\ t = 1 \ldots CT_{max} \qquad (10)$$

where $N(t)$ denotes a number of entity pairs (v_i, v_j) whose cardinality measure $CT_{ij} \geq t, t \in \{1 \ldots CT_{max}\}$. Figure 3 presents the density function derived from the Terrorist dataset [10], where $CT_{max} = 113$ (and the magnified presentation of $D(t), t \in \{7, 113\}$ is included herein for better interpretation).

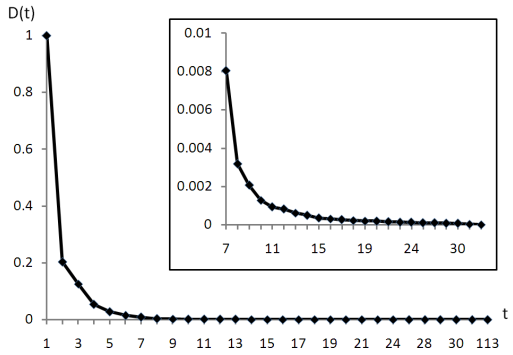

Fig. 3. Example of density function derived from Terrorist dataset.

Another useful functional concept is the average uniqueness degree. This is calculated from each group of entity pairs with a specific cardinality value. Let $\overline{UQ} : \{1, \ldots, CT_{max}\} \rightarrow [0, 1]$ be the averaged uniqueness function, which is specified by

$$\overline{UQ}(t) = \frac{\sum_{\forall (v_i, v_j), CT_{ij} = t} \overline{UQ}_{ij}}{M(t)},\ t = 1 \ldots CT_{max} \qquad (11)$$

where $\overline{UQ}_{ij}$ and $M(t)$ denote the average value amongst members in the uniqueness set UQ_{ij} (of the entity pair (v_i, v_j)), and a number of entity pairs whose cardinality measure $CT_{ij} = t, t \in \{1 \ldots CT_{max}\}$, respectively. In particular, Figure 4 shows the corresponding function of the Terrorist dataset.

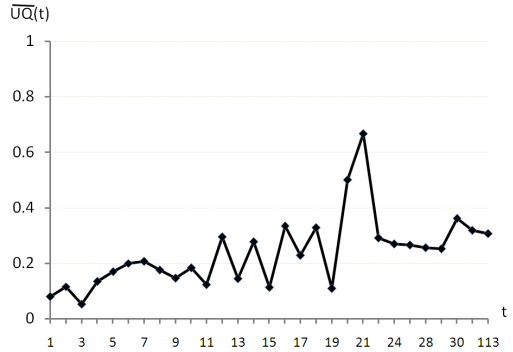

Fig. 4. The average uniqueness function derived from Terrorist dataset.

With these link-based functions, the following set of heuristics can be articulated especially to help data analysts to assess a proper value of $\beta \in \{1 \ldots CT_{max}\}$:

- The density value at β, $D(\beta)$, should not be too large such that the exclusion of high cardinality measure $(CT > \beta)$ is rational; intuitively, $D(\beta) \leq 0.1$.
- A particular value t is considered as a candidate for β if $\overline{UQ}(t) > \overline{UQ}(t+1)$, where $t, t+1 \in \{1 \ldots CT_{max}\}$. In other words, any higher cardinality measure that leads to lower link quality (in term of uniqueness) should be excluded.
- If two or more values satisfy the previous requirements, the one with the highest density value $D(t)$ is preferred.

This semi-supervised method is effective to assist analysts to design an appropriate stress function, 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.

V. Application and Performance Evaluation

A. Application to Alias Detection

Discovering duplicates and similar objects is a major subject in the fields of information retrieval, database and intelligence data analysis. Initial attempts to resolve the problem of aliases rely on text-based comparison, which suffer from their requirements of domain-specific rules/grammars for comparison, and of a typically vast amount of computational time and space [6]. Accordingly, the *link analysis approach* to this ambiguation problem has been put forward to underpin the accountability for unstructured information. It has proven effective for a wide range of domains, including personal name resolution in publication databases [12], webpage similarity [14], personal name resolution in emails [15], alias detection in spam emails and terrorism-related datasets [3], [10]. In practice, disclosing an alias pair in a link network $G(V, E)$ 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 constitute the actual alias pair.

This section presents an application of the OWA aggregation model to detecting alias pairs of references, each referring to the same real-world entity. Particularly, its performance is empirically evaluated, against state-of-art link-based algorithms, over a variety of data collections.

B. Experimented Datasets

The performance and applicability of the proposed approach is evaluated over the following distinct datasets: Terrorist [10], DBLP [12] and EmailThread [15]. 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 5 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 the bibliographical database. In this link network of publication information, each node represents a reference name

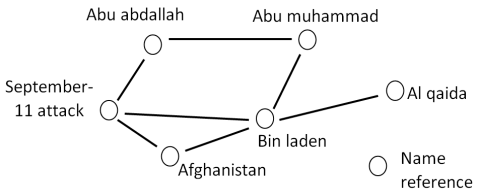

Fig. 5. An example of Terrorist dataset.

of an author and a link denotes the fact that two names appear as the co-authors of a paper (or papers). EmailThread is the subset of the email graphs used in [15] for the task of detecting email threads (i.e. similar email messages). The original dataset was extracted from email accounts in the Enron corpus. Objects in this dataset are references of email messages and keywords appearing in email subject fields. In particular, a link is only formed between a message reference and a particular keyword, indicating that the message's subject contains this keyword. Table I summarizes the number of links, objects and alias pairs included in these datasets.

TABLE I
DATASET DETAILS (NUMBER OF OBJECTS, LINKS AND ALIAS PAIRS).

Dataset	Objects	Links	Alias pairs
Terrorist	4088	5581	919
DBLP	2796	8157	23
EmailThread	4319	8955	106

C. Performance Evaluation

1) Efficiency of Semi-Supervised Method: Initially, it is important to examine the effectiveness of the semi-supervised method for modeling a stress function (i.e. selecting β value). By following the heuristics previously prescribed, appropriate β values for Terrorist, DBLP and EmailThread are 7, 27 and 5 (with corresponding CT_{max} of 113, 94 and 9), respectively. Figure 6 presents precision measures of the aggregation model, over EmailThread dataset, with two different stress functions (i.e. Stress1 and Stress2, see Figure 2) and two different β values (i.e. max and β denoting the maximum cardinality CT_{max} in the studied dataset and the value achieved from the semi-supervised method, respectively). Note that around 150 entity pairs with highest similarity values are included in this assessment.

These results suggest that the β values acquired through the semi-supervised process are more effective than those subjectively selected by an expert (e.g. CT_{max}). Similar observations are obtained with Terrorist and DBLP datasets, but they are not included here due to space limitation.

2) Aggregation Method vs. Other Link-based Measures: The performance of the aggregation model is assessed here, against three state-of-art link-based methods (Connected-Triple (CNT), SimRank (SR) and PageSim (PS)). Particularly, the Connected-Triple approach [12] estimates the similarity degree based solely on the cardinality measure. This neighbor-oriented intuition is extended through an iterative refinement of SimRank model [11] to find similar scientific papers through their citation relations. In a different domain, PageSim [14] was developed to capture similar web pages based on associations via their hyperlinks. Note that this algorithm explicitly uses the page ranking scheme, PageRank [5], of the Google search engine.

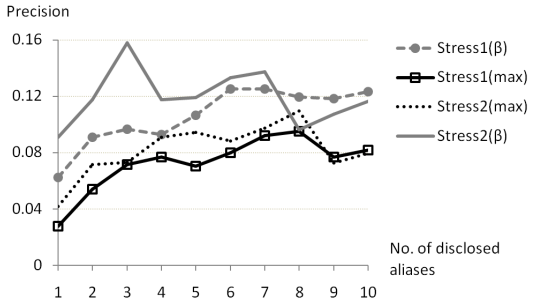

Fig. 6. Precisions of the aggregation model with different stress functions and β values, over EmailThread dataset.

Table II compares the number of disclosed alias pairs successfully detected by each method, where K denotes the number of entity pairs with highest similarity measures. These results suggest that the OWA aggregation models (i.e. $Stress1$ and $Stress2$) usually performs better than the rest over Terrorist and DBLP datasets, while being competitive to other more complex methods with EmailThread data.

TABLE II
NUMBER OF ALIAS PAIRS DISCLOSED BY EACH METHOD.

K	Stress1(β)	Stress2(β)	CNT	SR	PS
Terrorist					
200	43	9	1	0	7
400	81	66	5	0	36
600	117	107	74	1	63
800	150	146	77	1	79
1000	182	181	83	2	92
DBLP					
100	4	4	1	0	1
200	5	5	1	1	1
300	5	5	1	2	1
400	6	5	1	2	2
500	9	10	3	3	4
EmailThread					
100	12	12	3	13	6
200	19	21	10	13	18
300	27	29	10	36	27
400	33	33	23	36	35
500	36	40	23	47	38

In addition, Table III presents the recall measure as the number of alias pairs discovered within a whole dataset (i.e. K = the number of entity pairs identified with non-zero similarity value) by $Stress1$ and PS methods (note that the results of SR and PS are identical, while those of CNT and $Stress2$ are identical to that of $Stress1$). Accordingly, the performances of $Stress1$, $Stress2$ and other link-based measures are usually analogous, except in the case of Terrorist dataset where the recalls of $Stress1$ and $Stress2$ techniques are lower than those of the PS and SR. However, the number of false positives generated by the OWA aggregation models is far less than those by the others.

TABLE III
NUMBER OF ALIAS PAIRS DISCLOSED FROM ALL ENTITY PAIRS WITH NON-ZERO SIMILARITY VALUE.

Dataset	Stress1		PS	
	Alias	Non-Zero	Alias	Non-Zero
Terrorist	366	81,985	468	708,613
DBLP	21	20,782	21	23,163
EmailThread	70	1,915	70	10,196

3) 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 both OWA aggregation ($Stress$) and CT methods 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 II 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 (t is 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 II). As a result, the overall time complexity of PageSim method is $O(n^2m^{2r} + nmt)$.

Hence, the OWA aggregation 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.

VI. CONCLUSION

This paper has presented a new OWA aggregation model that can be exploited to derive link-based similarity measures, efficient for detecting aliases in a variety of real-world data collections. The proposed method is based on the OWA aggregation of multiple link measures, in which a stress function is particularly exploited to determine aggregation behavior. As a uniformly efficient stress function for various problems is generally difficult to obtain, a semi-supervised method is thus introduced such that data analysts can acquire an appropriate data-dependent function. In spite of successes reported herein, it is important to further generalize the proposed methodology with respect to other link datasets, and also to investigate the use of different types of stress function.

ACKNOWLEDGMENT

This work is sponsored by the UK EPSRC grant EP/D057086. The authors are grateful to the members of the project team for their contribution, but will take full responsibility for the views expressed in this paper.

REFERENCES

[1] G. Beliakov, A. Pradera, and T. Calvo. *Aggregation Functions: A Guide for Practitioners*. Springer, 2007.
[2] T. Boongoen and Q. Shen. Clus-DOWA: A New Dependent OWA Operator. In *Proceedings of IEEE International Conference on Fuzzy Sets and Systems*, pages 1057–1063, 2008.
[3] T. Boongoen and Q. Shen. Detecting false identity through behavioural patterns. In *Proceedings of Int. Crime Science Conference*, 2008.
[4] G. Bordogna, M. Fedrizzi, and G. Pasi. A linguistic modeling of consensus in group decision making based on OWA operators. *IEEE Transactions on Systems, Man and Cybernetics - Part A*, 27(1):126–133, 1997.
[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
[6] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. A. Ribeiro-Neto, and N. Ziviani. Link-based similarity measures for the classification of web documents. *Journal of American Society for Information Science and Technology*, 57(2):208–221, 2006.
[7] D. Filev and R. R. Yager. On the issue of obtaining OWA operator weights. *Fuzzy Sets and Systems*, 94(2):157–169, 1998.
[8] R. Fuller. On obtaining OWA operator weights: A short survey of recent developments. In *Proceedings of IEEE International Conference on Computational Cybernetics*, pages 241–244, 2007.
[9] L. Getoor and C. P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12, 2005.
[10] P. Hsiung, A. Moore, D. Neill, and J. Schneider. Alias detection in link datasets. In *Proceedings of International Conference on Intelligence Analysis*, 2005.
[11] G. Jeh and J. Widom. Simrank: A measure of structural-context similarity. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 538–543, 2002.
[12] S. Klink, P. Reuther, A. Weber, B. Walter, and M. Ley. Analysing social networks within bibliographical data. In *Proceedings of International Confonference on Database and Expert Systems Applications*, pages 234–243, 2006.
[13] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.
[14] Z. Lin, I. King, and M. R. Lyu. Pagesim: A novel link-based similarity measure for the world wide web. In *Proceedings of International Conference on Web Intelligence*, pages 687–693, 2006.
[15] E. Minkov, W. W. Cohen, and A. Y. Ng. Contextual search and name disambiguation in email using graphs. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34, 2006.
[16] H. B. Mitchell and D. D. Estrakh. A modified OWA operator and its use in lossless DPCM image compression. *International Journal of Uncertain Fuzziness, Knowledge Based Systems*, 5:429–436, 1997.
[17] M. O'Hagan. Aggregating template rule antecedents in real-time expert systems with fuzzy set logic. In *Proceedings of Annual IEEE Conference on Signals, Systems, and Computers*, pages 681–689, 1988.
[18] V. Torra and Y. Narukawa. *Modeling Decisions: Aggregation Operators and Information Fusion*. Springer, 2007.
[19] A. G. Wang, H. Atabakhsh, T. Petersen, and H. Chen. Discovering identity problems: A case study. In *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, pages 368–373, 2005.
[20] J. W. Wang, J. R. Chang, and C. H. Cheng. Flexible fuzzy OWA querying method for hemodialysis database. *Soft Computing*, 10(11):1031–1042, 2006.
[21] R. R. Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18:183–190, 1988.
[22] R. R. Yager. Connectives and quantifiers in fuzzy sets. *Fuzzy Sets and Systems*, 40:39–75, 1991.
[23] R. R. Yager. Quantifier guided aggregation using OWA operators. *International Journal of Intelligent Systems*, 11:49–73, 1996.
[24] R. R. Yager. Using stress functions to obtain OWA operators. *IEEE Transactions on Fuzzy Systems*, 15(6):1122 – 1129, 2007.
[25] R. R. Yager, L. S. Goldstein, and E. Mendels. Fuzmar: an approach to aggregating market research data based on fuzzy reasoning. *Fuzzy Sets and Systems*, 68(1):1–11, 1994.