Extensions to Pairwise Similarity Calculation of Information Networks

Yuanzhe Cai and Sharma Chakravarthy

Technical Report CSE-4-2010
Extensions to Pairwise Similarity Calculation of Information Networks

Yuanzhe Cai and Sharma Chakravarthy

CSE Department and Information Technology Laboratory
The University of Texas at Arlington, Arlington, TX 76019, USA
yuanzhe.cai@mavs.uta.edu, sharma@cse.uta.edu

Abstract. We focus on extensions to the pairwise similarity calculation of information networks. By considering both in- and out-link relationships, we propose Additive- and Multiplicative-SimRank to calculate the similarity score. Then we discuss the loop/cycles problem of information networks and propose a method to address this problem. Our extensive experimental results conducted on eight food web data sets show that our approach performs significantly better than earlier approaches.

1 INTRODUCTION

In order to study the patterns and processes of information systems, computing pairwise similarity in an information network is a fundamental problem. Food web1, a kind of information network, represents the predator-prey relationship between species within an ecosystem. Consider the following example from the food web.

Example 1. (Motivation). The dodos2 lived peacefully on Mauritius Island for several hundred years. Because of poaching by the humans and killing by the animals (such as pigs, rats and cats), that have been introduced into the island by sailors, the Dodo bird died off extremely quickly. About 1681, the last Dodo bird died. After about three hundred years, in 1973, Tambalacoque, also called dodo tree, was dying out. There are only about 13 trees living in the island. Scientists found that the dodo tree’s seed should pass through the digestive system of dodo before they germinated. Therefore, in order to aid the seed in germination, scientists used turkeys to erode the nutshell of the dodo tree seed. In this case, the humans saved the dodo tree, but the turkey, similar to the cats, rats and pigs, which have been introduced into this island may also spoil the balance of the ecosystem. Some failed examples, such as Austrian Rabbit and Xisha Islands’ cat, are also alarming. Therefore, there is an interesting question, “If one species get extinct in an ecosystem and we want to introduce a new species into this ecosystem to keep the balance, what kinds of species should we introduce?”

The answer is that we should introduce the species that has “similar food habit” in this ecosystem. In this example, turkey and dodo bird are very similar, because they both eat similar foods, such as the seed of the dodo tree and they both have similar natural enemies. In this example, the prey relationship and is-preyed relationship are used to define the similarity score between two species. Based on this observation, a number of approaches have been proposed to quantify similarity between species in a food web. The most widely used approaches in the former research work [10] [11] [14] are the Jaccardian similarity functions. The intuition behind Jaccardian similarity function is that two species are similar, if they share many similar food web neighbors and the total number of their neighbors is less. The $S_{jaccard}(a, b)$ [10] equation is shown below:

$$S_{jaccard}(a, b) = \left( \frac{|n(a) \cap n(b)|}{|n(a) \cup n(b)|} \right)$$ (1)

1http://en.wikipedia.org/wiki/Food_web
2http://en.wikipedia.org/wiki/Dodo
where \( n(a) \) and \( n(b) \) is the neighbors of species \( a \) and \( b \). \( |n(a) \cap n(b)| \) is the total number of prey and predator species that species \( a \) and \( b \) have in common and \( |n(a) \cup n(b)| \) is the total number of prey and predators of species \( a \) and \( b \).

Fig. 1: Segment of CYPWET data set [3]

However, earlier researches only consider the direct relationship in the information network. Considering the example in Figure 1, we want to calculate the similarity between fishing spider and crayfish. However, according to the direct relationship, shown in Table 1, the similarity between these two species is zero, although these two species have some relationships from theirs indirect predator – gruiformes. This example shows that when we consider the similarity between two species in the food web, we also need to consider the indirect relationship between species of other types related to them. This problem is addressed by the SimRank algorithm [4] in which the similarity between two objects is recursively defined as the average similarity between two objects. However, this similarity definition only considers the one directional relationship for the information network. In Figure 1, SimRank only considers the is-preyed relationship (their predators), but the species’ similar preies also contribute to their similarity relationship.

The other major problem is that many real-life information networks contain cycles. Food web, for example, includes many cannibals that create self-loops and cycles. For example, in figure 1, salamander is a cannibal and preys the other salamanders for food. These loops in the food web also influence the value of similarity scores.

The main contributions of this are:

• Based on relationship of topical structures in the information networks, two similarity algorithms, Additive-SimRank and Multiplicative-SimRank, are proposed to address this problem. We also prove that the proposed algorithms converge by theoretical analysis.

• We discuss the loops and cycles problem in the information networks and propose a method to handle them.

• Extensive experiments are conducted to evaluate the accuracy of the proposed algorithms. Additive-SimRank is shown to have higher accuracy as compared with other methods.

Roadmap: The rest of this paper is organized as follows: We introduce the related work in section 2 and define the graph model in section 3. SimRank is described in section 4. Two similarity measures for information network and the cycle problem are discussed in section 5. Our experimental analysis is reported in section 6 and conclusions are in section 7.

2 RELATED WORK

We categorize existing work related to our study into three classes: species aggregation, link-Based similarity calculation, and random walk on graph.

Species aggregation: Setting a new criterion for searching community food web data, Martineze [10] [11] was the first researcher to systematically analyze the effects of variable species aggregation on the network structure of food webs. There are different indices used to quantify
similarity between objects. The Jaccard index is probably the best known and widely used in food web research [10] [11] [14]. Martineze used an Additive-Jaccard index to determine the similarity between species in Little Rock Lake and then used the average-linkage-cluster to aggregate taxonomies. However, these methods do not consider the potential relationship between each species.

**Link-based similarity calculation:** The earliest research works for similarity calculation based on link analysis focus on the citation patterns of scientific papers. The most common measures are co-citation [12] and co-coupling [6]. Co-citation indicates that if two documents are often cited together by other documents, they may have the same topic. However, all these methods compute similarity only by considering their immediate neighbors. In contrast, SimRank [4] considers the entire relationship graph to determine similarity between two nodes. Because of the high time complexity ($O(n^4)$) of this approach, many papers [2] [13] [1] [8] have focused on performance improvement, but a few have focused on the accuracy improvement of SimRank. In this paper, our focus is on extending SimRank approach by considering bidirectional relationships and cycles to improve the accuracy of the original SimRank approach.

**Random walks on graphs:** Theoretical basis of our work uses hit times for two surfers walking randomly on the graph. We mainly refer to research about expected f-meeting distance theory [4]. Other researches, such as random walk theory [9] and Markov Model [7], also help understand our research.

### 3 GRAPH MODEL

The food web data can be represented as a directed graph, $G(V, E)$, which consists of a set of nodes $V$ representing species and a set of directed edges $E$ representing the relationships between species. For example, Figure 1 is a relationship graph that describes the predatory relationship in the marshes and sloughs. In this graph, a directed edge $< p, q >$ from species $p$ to species $q$ corresponds to a predator relationship. $I(v)$ denotes the set of predators preying on species $v$, which is also the in-link neighbors of species $v$ and $O(v)$ denotes the set preyed-by species $v$, which is also the out-link neighbors of species $v$.

### 4 OVERVIEW OF SimRank

SimRank [4] is a method for measuring link-based similarity between objects in a graph that models the object-to-object relationships in a particular domain. The intuition behind SimRank score is that two objects are similar if they link to similar objects. This intuition also indicates that SimRank calculation needs to be recursive.

Below, we present the formula to compute SimRank. Given a graph $G(V, E)$ consisting of a set of nodes $V$ and a set of links $E$, the SimRank similarity between objects $a$ and $b$, denoted as $S(a, b)$, is computed, recursively, as follows:

$$S(a, b) = \begin{cases} 
\frac{1}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} S(I_i(a), I_j(b)) & \text{if } (a = b) \\
\frac{c}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} S(I_i(a), I_j(b)) & \text{if } (a \neq b)
\end{cases} \quad (2)$$

where $c$ is a constant decay factor, $0 < c < 1$; $I(a)$ is the set of in-neighbor nodes of $a$ and $I_i(a)$ is the $i^{th}$ in-neighbor node of $a$. $|I(a)|$ is the number of neighbors of node $a$. In case that $I(a)$ or $I(b)$ is an empty set, $S(a, b)$ is defined as zero.

A solution to SimRank equation (2) can be reached by iteration to a fixed-point. For each iteration $k$, let $S_k(., .)$ be an iteration similarity function and $S_k(a, b)$ be the iterative similarity score of pair $(a, b)$ on iteration $k$. The iteration process is started with $S_0 (S_0(a, b) = \begin{cases} 
0 & \text{if } (a \neq b) \\
1 & \text{if } (a = b)
\end{cases}$.
To calculate \( S_{k+1}(a, b) \) from \( S_k(a, b) \), we use the following equation:

\[
S_{k+1}(a, b) = c \frac{|I(a)||I(b)|}{|I(a)||I(b)|} \sum_{i=1}^{||I(a)||} \sum_{j=1}^{||I(b)||} S_k(I_i(a), I_j(b))
\]

In equation (3), \( 1/|I(a)| \) is a single step probability of walking from node \( a \) to a node in \( I(a) \). Therefore we can use Backward Transfer Probability Matrix (BT PageRank) to capture the single step probability in a Markov Chain. Thus, SimRank algorithm can be described by matrix calculation. \( S_0 = E \), where \( E \) is an identity matrix. Equation (3) can be rewritten as:

\[
S_k(a, b) = c \sum_{i=1}^{||I(a)||} \sum_{j=1}^{||I(b)||} BT_{aI_i(a)} BT_{bI_j(b)} S_{k-1}(I_i(a), I_j(b))
\]

Although the convergence of iterative SimRank algorithm can be guaranteed in theory, practical computation uses a tolerance factor \( \varepsilon \) to control the number of iterations such that a finite number of iterations are performed. It is recommend to set \( \varepsilon = 0.001 \), the same as in PageRank. Specifically, the terminating condition of the iteration is as follows:

\[
\max(|S_k(a, b) - S_{k-1}(a, b)|/|S_{k-1}(a, b)|) \leq \varepsilon
\]

It indicates that the iteration stops if the maximal change rate of similarity value between two iterations for all node pairs is smaller than the threshold \( \varepsilon \).

5 EXTENDING THE SIMILARITY MEASURE

In this section, we first describe our analysis of the information network. Then, we describe our topological similarity definition on the network. Finally, we discuss the loops problem on the network.

5.1 Topological Similarity

If we want to compare the similarity between dodo and turkey on the Mauritius Island, we need to answer the following questions:

1. Do dodo bird and turkey eat similar food? If turkey does not eat dodo tree’s seed, we do not need to introduce turkey into this ecosystem, because turkey doesn’t have the similar role as the dodo bird in this ecosystem.
2. Do dodo bird and turkey have similar natural enemies? If turkey does not have similar natural enemies as the dodo bird or do not have natural enemies, the dodo bird’s natural enemies may not find enough food and also become extinct; or turkeys may proliferate and break the biological balance.

Thus, we can identify two intuitions for defining the similarity for the food web.

**Intuition 1:** Two species are similar, if they are preyed by similar species.

**Intuition 2:** Two species are similar, if they prey similar species.

Let us look at Table 1 again. Surprisingly, SimRank doesn’t produce a similarity score for the pair “gruiforms”-“ducks”, although these two species have the same classification (avifauna) and prey the same species salamander. The problem for SimRank is that SimRank only considers is-preyed relationship on the food web, but the other important prey relationship is not considered for similarity calculation.
Considering both relationships for the food web, similarity score should combine the similarity from both relationships. Thus we can add is-preyed relationship similarity score and prey relationship similarity score together and use the parameter \( \gamma \) to adjust the contribution of these two relationships for the total score. We call this the additive method. Thus, we propose the following formula for calculating the similarity score:

\[
S(a, b) = \begin{cases} 
1 & \text{if } (a = b) \\
\gamma \frac{c}{|I(a)||O(b)|} \sum_{i=1}^{\lfloor I(a) \rfloor} \sum_{j=1}^{\lfloor I(b) \rfloor} S(I_i(a), I_j(b)) + (1 - \gamma) \frac{c}{|O(a)||O(b)|} \sum_{i=1}^{\lfloor O(a) \rfloor} \sum_{j=1}^{\lfloor O(b) \rfloor} S(O_i(a), O_j(b)) & \text{if } (a \neq b)
\end{cases}
\]

\( \)where, \( c \) is a constant decay factor, \( 0 < c < 1; I(a) \) is the set of predators of species \( a \) and \( I_i(a) \) is the \( i^{th} \) predators of \( a \), \( |I(a)| \) is the number of predator of node \( a \), \( O(a) \) is the set of prey of species \( a \) and \( O_i(a) \) is the \( i^{th} \) prey of \( a \), \( |O(a)| \) is the number of prey of node \( a \). \( \gamma \) is a constant parameter that use to adjust the different effect of the is-preyed and prey relationships, \( 0 \leq \gamma \leq 1 \).

On the other hand, another way to extend SimRank is that we can multiply the is-preyed and prey relationship similarities. This product score can also describe the relationship similarity score. This method is called as the multiplicative method. Then, we have the following formula to calculate the similarity score.

\[
S(a, b) = \begin{cases} 
1 & \text{if } (a = b) \\
\gamma \frac{c}{|I(a)||O(b)|} \sum_{i=1}^{\lfloor I(a) \rfloor} \sum_{j=1}^{\lfloor I(b) \rfloor} S(I_i(a), I_j(b)) \times (1 - \gamma) \frac{c}{|O(a)||O(b)|} \sum_{i=1}^{\lfloor O(a) \rfloor} \sum_{j=1}^{\lfloor O(b) \rfloor} S(O_i(a), O_j(b)) & \text{if } (a \neq b)
\end{cases}
\]

where parameter definitions are the same as that of the Additive case.

**Algorithm 1 Additive-SimRank**

**Require:**
- Decay Factor, \( c \);
- Tolerance Factor, \( \epsilon \);
- Forward Transfer Probability Matrix \( FT \) (the forward probability of moving from state \( i \) to state \( j \) in one step);
- Backward Transfer Probability Matrix \( BT \) (the backward probability of moving from state \( i \) to their \( j \));

**Ensure:**
- Similarity Matrix, \( S_k \);
1: \( k \leftarrow 1 \);
2: \( S_0 \leftarrow \text{identity} \);
3: while(\( \text{Max}(|S_k(a, b) - S_{k-1}(a, b)|)/|S_{k-1}(a, b)|) > \epsilon \))
4: \( k \leftarrow k+1 \);
5: \( S_{k-1} \leftarrow S_k \);
6: for each element \( S_k(a, b) \)
7: \( S_k(a, b) \leftarrow \gamma c \sum_{i=1}^{\lfloor I(a) \rfloor} \sum_{j=1}^{\lfloor I(b) \rfloor} BT_{a_i(a)} BT_{b_i(b)} S_{k-1}(I_i(a), I_j(b)) + (1 - \gamma) c \sum_{i=1}^{\lfloor I(a) \rfloor} \sum_{j=1}^{\lfloor I(b) \rfloor} FT_{a_i(a)} FT_{b_i(b)} S_{k-1}(I_i(a), I_j(b)) \);
8: end for;
9: end while;
10: return \( S_k \);

Algorithm 1 outlines Additive-SimRank computation. It takes in 4 arguments. The first two arguments inherit from the original SimRank algorithm: the decay factor \( c \) gives the rate of decay as similarity flows across edges in a graph and tolerance factor \( \gamma \) is to control the number of iterations as discussed in section 4. The next parameter is Forward Transfer Probability Matrix \( FT \). As we can see from equation 6, \( 1/|O(a)| \) is a single step probability of walking from node \( a \) to a node in \( O(a) \). Thus, we use the Forward Transfer Probability Matrix \( FT \) [7] to calculate the similarity score in our algorithm. On the food web, \( FT \) matrix is the transfer matrix of prey
relationship. The last parameter is Backward Transfer Probability Matrix $BT$. As we can also see from equation $6$, $1/|I(a)|$ is a single step probability of walking from node $a$ to a node in $I(a)$. Thus, we use the Backward Transfer Probability Matrix $BT$ [7] to calculate the similarity score in our algorithm. On the food web, $BT$ matrix is the transfer matrix of is-preyed relationship.

Additive-SimRank algorithm first initializes variables (lines 1-2). In line 4, the algorithm will stop if the ending condition, equation 5, will be satisfied. The algorithm then uses Equation 6 to calculate the similarity score. Although the worst time and space complexity of Additive-SimRank is the same as the SimRank, its accuracy of Additive-SimRank is higher than original SimRank as it considers the both relationship of the graph.

The Multiplicative-SimRank algorithm is the same as the previous algorithm except for step 7 where Equation 7 is used. The theoretical foundations of Additive-SimRank and Multiplicative-SimRank are discussed below.

Forward and Backward Random Walk Model: Since $BT$ and $FT$ in algorithm 1 (and its counterpart for multiplicative-SimRank) can be considered as a single step backward and forward transfer matrix of a Markov Chain, the iteration similarity calculation process of equations 6 and 7 can be explained using two random surfers walking forward and backward. Two surfers start from two nodes on the graph and they walk from one node to the other nodes step by step. In each step, they will walk one step backward or forward, respectively, and calculate the meeting possibility for these two surfers. The final result of these two methods can be translated into the possibility of two random surfers meeting with each other by considering both forward and backward random walking. For equations 6 and 7, we use different methods to combine these meeting possibilities for two random surfers meeting with each other by considering both forward and backward random walking. For equations 6 and 7, we add these meeting possibilities of forward and backward walking and use $\gamma$ to adjust the proportion of these backward and forward meeting possibilities. In equation 7, we directly multiply the forward and backward meeting score. Since SimRank only considers backward random walk, it is a special case of our method. In equation 6, if $\gamma$ is set to 1, the equation is the same as the SimRank function. The multiplicative form distinguishes the roles of predator and prey for each species, and requires a high similarity in both roles to achieve an overall high score. On the other hand, unlike the multiplicative form, additive form uses the parameter $\gamma$ to adjust the weight that need to be associated with predator and prey for each species.

We also give the convergence proof of Additive-SimRank and Multiplicative-SimRank as follow.

Lemma 1. Let $AR_k(a,b)$ be the score of $k^{th}$ iteration.

\[
\text{Then, } AR_{k+1}(a,b) - AR_k(a,b) \leq 0
\]  

Proof. If $a = b$, then $AR_{k+1}(a,b) = 1$, $AR_k(a,b) = 1$ by definition, $AR_{k+1}(a,b) - AR_k(a,b) = 0$ and thus (12) holds.

In the same way, if $I(a), I(b) = \emptyset$ and $O(a), O(b) = \emptyset$, then by definition, $AR_{k+1}(a,b) = AR_k(a,b) = 0$, $AR_{k+1}(a,b) - AR_k(a,b) = 0$ and thus (12) holds.

Induction Base Step: Let us prove that (13) holds for $k = 0$, i.e. that for every two nodes $a, b$: $AR_1(a,b) - AR_0(a,b) \geq 0$.

If $a \neq b$, $AR_0(a,b) = 0$. $AR_1(a,b)$ is defined by the iterative equation (1) as follow.

\[
AR_1(a,b) - AR_0(a,b) = AR_1(a,b) \geq 0
\]

Inductive Step: Provided that $AR_k(a,b) - AR_{k-1}(a,b) \geq 0$, let’s prove that (12) hold for $(k + 1)$ as well:

\[
AR_{k+1}(a,b) - AR_k(a,b) = (\gamma \sum_{i=1}^{[I(a)]} \sum_{j=1}^{[I(b)]} AR_k(I_i(a), I_j(b)) + (1-\gamma) \sum_{i=1}^{[O(a)]} \sum_{j=1}^{[O(b)]} AR_k(O_i(a), O_j(b))) - (\gamma \sum_{i=1}^{[O(a)]} \sum_{j=1}^{[O(b)]} AR_{k-1}(I_i(a), I_j(b)) + (1-\gamma) \sum_{i=1}^{[O(a)]} \sum_{j=1}^{[O(b)]} AR_{k-1}(O_i(a), O_j(b)))
\]
herpetofauna. In fact, “fishing spider” and “salamander” are not in the same classification. Similarly, fishing spider and apple snail are classified as macro invertebrates but salamander is classified as salamander. However, in the biological field, fishing spider and apple snail are quite common in the food web. Each other. Table 2 shows the number of cycles in the real world food web [3]. As we can see, cycles loops in the network. For example, food web contains frequent cannibalism that induces loops (e.g., salamander in Figure 1). In the dry season, 14% of salamanders’ food comes from killing other salamanders. Another example is of steatoda spiders and latrodectus spiders. These two spiders eat each other. Table 2 shows the number of cycles in the real world food web [3]. As we can see, cycles are quite common in the food web.

\[ ((1 - \gamma) \frac{c}{|V(a)||V(b)|} \sum_{i=1}^{|V(a)|} \sum_{j=1}^{|V(b)|} (AR_k(I_i(a), I_j(b)) - AR_{k-1}(I_i(a), I_j(b))) + \]
\[ (1 - \gamma) \frac{c}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} (AR_k(O_i(a), O_j(b)) - AR_{k-1}(O_i(a), O_j(b))) \]
\[ = \gamma c + (1 - \gamma)c = c \leq 1 \]

Thus, \( AR_k(a, b) \) will converge to a fixed value.

5.2 Dealing with Loops in the network

The other problem of some information networks is that there could be a number of cycles or loops in the network. For example, food web contains frequent cannibalism that induces loops (e.g., salamander in Figure 1). In the dry season, 14% of salamanders’ food comes from killing other salamanders. Another example is of steatoda spiders and latrodectus spiders. These two spiders eat each other. Table 2 shows the number of cycles in the real world food web [3]. As we can see, cycles are quite common in the food web.

Fig. 2: Representative graph with a loop/cycle

However, these cycles on the food web graph will affect the species’ similarity score. Let us look at the similarity score between two species “fishing spider”, “salamander” and “fishing spider”, “apple snail”. Table 3 tabulates these similarity score for figure 1. As we can see \( S(\text{fishing spider, salamander}) \) is slightly higher than \( S(\text{fishing spider, apple snail}) \). However, in the biological field, fishing spider and apple snail are classified as macro invertebrates but salamander is classified as herpetofauna. In fact, “fishing spider” and “salamander” are not in the same classification. Similarly,

<table>
<thead>
<tr>
<th>Data set</th>
<th>Vertex</th>
<th>Edge</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYPWET</td>
<td>68</td>
<td>554</td>
<td>15</td>
</tr>
<tr>
<td>CYPDRY</td>
<td>68</td>
<td>545</td>
<td>15</td>
</tr>
<tr>
<td>BAYWET</td>
<td>125</td>
<td>1969</td>
<td>21</td>
</tr>
<tr>
<td>BAYDRY</td>
<td>125</td>
<td>1938</td>
<td>21</td>
</tr>
<tr>
<td>MANGWET</td>
<td>94</td>
<td>1389</td>
<td>8</td>
</tr>
<tr>
<td>MANGDRY</td>
<td>94</td>
<td>1440</td>
<td>8</td>
</tr>
<tr>
<td>GRAMWET</td>
<td>66</td>
<td>793</td>
<td>10</td>
</tr>
<tr>
<td>GRAMDRY</td>
<td>66</td>
<td>793</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 2: Statistics in Food Web Data sets
other information networks, such as the web page graphs and paper citation graphs, also has cycles. For example, in the citation graph, the same author can write two papers that are cross-referenced with each other. We can also actually prove that the following theorem for the similarity calculation in the presence of loops in a graph.

\[
S(s_1, s_n) = \sum_{i=1}^{n_1} P[i] + c^2 \sum_{i=1}^{n_2} P[i] + ... + c^k \sum_{i=1}^{n_k} P[i] + ...
\]

According to the definition of \(G\), \(length(l) = length(q) = k\). Then, \(S(s_1, s_n) = c^k\) and \(S(s_n, s_m) = 0\).

**Proof.** (i) According to the SimRank definition [4], \(S(s_1, s_n) = \sum_{i=1}^{n_1} P[i] \cdot c^i\) and \(S(s_n, s_m) = c^k\) for all \(i\) and \(j\).

This theorem provides us two insights about SimRank scores and why they are not intuitively right for the networks that contain loops. First, \(s_1\) is at the bottom of food web in Figure 2 and in normal cases it is the primary species, such as periphyton, unicellular, and so on. However, \(s_n\) is the top consumer, such as bobcat, panther and so forth. However, according to theorem 1, these two species \(s_1\) and \(s_n\) have a greater similarity between each other. Secondly, \(s_n\) is another species in the cycle. In this food web graph, this species is also the top level consumer. However, according to theorem 1, the pair \(s_1\) and \(s_n\) have higher similarity score than the pair \(s_m\) and \(s_n\). This implies that bobcat and periphyton are more similar than bobcat and panther. That does not match with our intuition.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.06</td>
<td>0.02</td>
<td>0.01</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>2</td>
<td>0.09</td>
<td>0.09</td>
<td>0.065</td>
<td>0.065</td>
<td>0.065</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.09</td>
<td>1</td>
<td>0.02</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>4</td>
<td>0.002</td>
<td>0.06</td>
<td>0.002</td>
<td>1</td>
<td>0.21</td>
<td>0.08</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>0.03</td>
<td>0.18</td>
<td>0.31</td>
<td>1</td>
<td>0.21</td>
<td>0.154</td>
<td>0.154</td>
</tr>
<tr>
<td>6</td>
<td>0.002</td>
<td>0.06</td>
<td>0.002</td>
<td>0.08</td>
<td>0.21</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.065</td>
<td>0.065</td>
</tr>
</tbody>
</table>

Table 3: Additive-SimRank results for figure 1(with cycles)(\(\gamma = 0.75, c = 0.8\))

Table 4: Additive-SimRank results for figure 1(no cycles)(\(\gamma = 0.75, c = 0.8\))
6 EXPERIMENTAL EVALUATION

Data Sets: Our experiments use the data sets shown in Table 2. Please refer to [3] for details regarding these data sets. Before we calculate the similarity score, we delete all the cycle in these data sets. These eight data sets come from four areas. CYPWET and CYPDRY data sets are collected from 295,000 hectare wetlands of the big cypress natural preserve in southwest Florida. BAYWET and BAYDRY data sets are collected from a triangular, tropical lagoon/bay. MANGWET and MANGDRY data sets are from the huge mangrove belt along the seaward edge of the Everglades. GRAMWET and GRAMDRY data sets are from the historical Everglades system. In each area, the food web data is collected for different seasons. For example, CYPWET indicates that this data set is collected in wet season and CYPDRY is for the dry season.

Table 5: Classification of for food web data sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>C.1</th>
<th>C.2</th>
<th>C.3</th>
<th>C.4</th>
<th>C.5</th>
<th>C.6</th>
<th>C.7</th>
<th>C.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYPWET</td>
<td>12</td>
<td>2</td>
<td>16</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>CYPDRY</td>
<td>12</td>
<td>2</td>
<td>16</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>BAYWET</td>
<td>14</td>
<td>12</td>
<td>26</td>
<td>4</td>
<td>48</td>
<td>3</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>BAYDRY</td>
<td>14</td>
<td>12</td>
<td>26</td>
<td>4</td>
<td>48</td>
<td>3</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>MANGWET</td>
<td>5</td>
<td>6</td>
<td>12</td>
<td>21</td>
<td>5</td>
<td>22</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>MANGDRY</td>
<td>5</td>
<td>6</td>
<td>12</td>
<td>21</td>
<td>5</td>
<td>22</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>GRAMWET</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>21</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>GRAMDRY</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>21</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

Note: The species in these data sets have been divided into eight classes by their different roles in ecosystem, such as primary producers, micro fauna, mammals, macro invertebrates, herpetofauna, fishes, detritus and avifauna, which are marked from C.1 to C.8.

Table 5 shows that these species data sets are manually divided into eight classes. These classes will be used as the standard/baseline to evaluate the accuracy of our algorithms.

All our experiments are conducted on a PC with a 3.0 GHz Intel Core 2 Duo Processor, 2GB memory, running windows XP Professional. All algorithms are implemented in Java.

6.1 Evaluation Metric

In our food web data sets, there are predefined class labels for these species. For a species on the food web, these algorithms will return a ranked list of relative species. For each species in the list, if this species' label is the same as species $s_1$, we think these two species are closely related and give a grade 2 (stress the related species); otherwise we associate grade 0. Then, we use the normalized discount cumulative gain (NDCG) [5] to evaluate the performance of this similarity ranking list. While evaluating a similarity ranking list, NDCG follows one principle. The lower ranking position of a species is less valuable for the researcher, because the researchers take great care about species more related to species $s_1$. According to this principle, NDCG score of a similarity ranking list at position n is calculated as follows. $N(n) = Z_n \sum_{j=1}^{n} \frac{r(j) - 1}{\log (1+j)}$, where $r(j)$ is the rating of the $j^{th}$ species in the similarity ranked list and the normalization constant $Z_n$ is chosen so that a prefect order gets NDCG value 1. For example, we will calculate the NDCG@10 score for the species “Living sediment” in data set CYPWET. Because for “living sediment” there is only one species in the micro fauna classification, $Z_n$ order is 2,0,0,0,0,0,0,0,0. We calculate NDCG within 10 related species for each species in each data set and get the average score to evaluate the validity of our experiments.
6.2 Experimental Results

Parameter Study: Two parameters, \( c \) and \( \gamma \), affect the accuracy of similarity scores directly. These two parameters are application dependent. We want to study the available parameters for the food web data.

First, we discuss the parameter \( \gamma \) for Additive-SimRank. This parameter is used to decide the importance of two relationships: is-preyed and prey for accuracy. In this experiment, we fix the damping factor \( c \) to 0.8 and vary \( \gamma \) from 0 to 1. Figure 6.2 shows that when \( \gamma \) is equal to 0.75, Additive-SimRank will receive the highest accuracy. Interestingly, is-preyed relationship is much more important to decide the species classification.

Second, we determine the damping factor \( c \) for these three link-based similarity algorithms. In this experiment, we fix \( \gamma = 0.75 \) and vary \( c \) from 0.05 to 0.95. In fact, the effect of damping factor \( c \) is not very obvious. Figure 6.2 shows that when \( c = 0.8, 0.1 \) and 0.95, Additive-SimRank, SimRank and Multiplicative-SimRank will receive the highest scores. Thus, for the rest of the experiments, \( \gamma \) is set to 0.75 for Additive-SimRank and \( c \) is set to 0.8, 0.1 and 0.95 for Additive-SimRank, SimRank and Multiplicative-SimRank, respectively.

![Fig. 3: Parameter \( \gamma \) for Additive-SimRank](image3.png)

![Fig. 4: Parameter \( c \)](image4.png)

Accuracy Analysis: In these experiments, we compare the accuracy among Multiplicative-Jaccard [14], Additive-Jaccard [14], Multiplicative-SimRank, Additive-SimRank and SimRank. Using the rule of additive and multiplicative methods, it is easy to design Multiplicative-Jaccard and Additive-Jaccard algorithm. Figure 5 shows the accuracy of these eight methods for food web data sets. We can see Multiplicative-Jaccard and Multiplicative-SimRank have the lowest accuracy. Because SimRank and Additive-SimRank consider the potential linkage information, these two algorithms are much better than Additive-Jaccard algorithm. Because Additive-SimRank considers both is-preyed and prey relationship, it reaches the best accuracy. Figure 6 plots the results of NDCG@1 to NDCG@19 for the each algorithm.

![Fig. 5: Segmentation of CYPWET data set [3]](image5.png)

![Fig. 6: NDCG@1 to NDCG@19](image6.png)
In Table 6, SimRank shows a better accuracy for two foodwebs: GRAMWET and GRAMDRY. This is primarily because we used the $\gamma$ for all of the data sets although it was derived for the CYPWET data set. When we use the correct $\gamma$ derived for this, the results are better for Additive-SimRank for these two as well.

Considering the case study, we analyze the top ten similar species for the species “Roots” in CYPWET food web. Because “Roots” is the primary producer, it only contains the is-preyed relationship. The result is shown in table 6. Because multiplicative method is the product of two relationships’ similarity score, Multiplicative-Jaccard and Multiplicative-SimRank can’t produce any similar species for “Roots”. On the other hand, Additive-Jaccard only considers the direct relationship, thus it only searches about eight species for “Roots” but no species are primary producers. The result of SimRank, containing seven primary producers, is also very good, but because Additive-SimRank considering both is-preyed and prey relationship, Additive-SimRank searches eight primary producers, which is slightly higher than SimRank.

Table 6: Case study for species “Roots”

<table>
<thead>
<tr>
<th>Multi.-Jaccard</th>
<th>Additive-Jaccard</th>
<th>Multi.-SimRank</th>
<th>Additive-SimRank</th>
<th>SimRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>Apple Snail</td>
<td>Null</td>
<td>Cypress Wood</td>
<td>Cypress Wood</td>
</tr>
<tr>
<td>Null</td>
<td>Crayfish</td>
<td>Null</td>
<td>RW Wood</td>
<td>RW Wood</td>
</tr>
<tr>
<td>Null</td>
<td>Prawn</td>
<td>Null</td>
<td>Vine Leaves</td>
<td>Vine Leaves</td>
</tr>
<tr>
<td>Null</td>
<td>Aquatic Invertebrates</td>
<td>Null</td>
<td>Cypress Leaves</td>
<td>Cypress Leaves</td>
</tr>
<tr>
<td>Null</td>
<td>Vertebrate Det.</td>
<td>Null</td>
<td>Vertebrate Det.</td>
<td>Epiphytes</td>
</tr>
<tr>
<td>Null</td>
<td>Refractory Det.</td>
<td>Null</td>
<td>Float. vegetation</td>
<td>Float. vegetation</td>
</tr>
<tr>
<td>Null</td>
<td>Liable Det.</td>
<td>Null</td>
<td>Macrophytes</td>
<td>Macrophytes</td>
</tr>
<tr>
<td>Null</td>
<td>null</td>
<td>Null</td>
<td>Phytoplankton</td>
<td>Living POC</td>
</tr>
<tr>
<td>Null</td>
<td>null</td>
<td>Null</td>
<td>Living POC</td>
<td>Living Sediment</td>
</tr>
</tbody>
</table>

7 CONCLUSIONS

In this paper, considering both prey (out-link) and is-preyed relationship (in-link) on the food web, we propose Additive- and Multiplicative-SimRank to calculate the similarity scores. Then, we also discuss the loop problem on the network and propose a method to address this problem. The experimental results conducted on eight food web data sets show that Additive-SimRank outperforms the other approaches with $\gamma$ equal to 0.75 (receives the highest score in the food web). In addition, our methods are also applicable for other information networks, such as paper citation network and web page network, that have similar characteristics.

References

8. D. Lizorkin, P. Velikhov, M. Grinev, and D. Turdakov. Accuracy estimate and optimization techniques for
simrank computation. The VLDB Journal The International Journal on Very Large Data Bases, 19(1):45 – 66,
February 2010.
10. N. D. Martinez. Artifacts or attributes? effects of resolution on the little rock lake food web. Ecological Mono-
13. X. Yin, J. Han, and P. S. Yu. Linkclus: efficient clustering via heterogeneous semantic links. In Proceedings of
14. P. Yodzis and K. O. Winemiller. In search of operational trophospecies in a tropical aquatic food web. Oikos,