

Preference-driven Similarity Join

Chuancong Gao
Simon Fraser University
cga31@sfu.ca

Jiannan Wang
Simon Fraser University
jnwang@sfu.ca

Jian Pei
Simon Fraser University
Huawei Technologies Co., Ltd.
jpei@cs.sfu.ca

Rui Li
Google Inc.
ruili@google.com

Yi Chang
Huawei Research America
yichang@acm.org

ABSTRACT

Similarity join, which can find similar objects (e.g., products, names, addresses) across different sources, is powerful in dealing with variety in big data, especially web data. Threshold-driven similarity join, which has been extensively studied in the past, assumes that a user is able to specify a similarity threshold, and then focuses on how to efficiently return the object pairs whose similarities pass the threshold. We argue that the assumption about a well set similarity threshold may not be valid for two reasons. The optimal thresholds for different similarity join tasks may vary a lot. Moreover, the end-to-end time spent on similarity join is likely to be dominated by a back-and-forth threshold-tuning process.

In response, we propose preference-driven similarity join. The key idea is to provide several *result set preferences*, rather than a range of thresholds, for a user to choose from. Intuitively, a result set preference can be considered as an objective function to capture a user’s preference on a similarity join result. Once a preference is chosen, we automatically compute the similarity join result optimizing the preference objective. As the proof of concept, we devise two useful preferences and propose a novel preference-driven similarity join framework coupled with effective optimization techniques. Our approaches are evaluated on four real-world web datasets from a diverse range of application scenarios. The experiments show that preference-driven similarity join can achieve high-quality results without a tedious threshold-tuning process.

1 INTRODUCTION

A key characteristic of big data is *variety*. Data (especially web data) often comes from different sources and the value of data can only be extracted by integrating various sources together. Similarity join,

This work is supported in part by the NSERC Discovery Grant program, the Canada Research Chair program, the NSERC Strategic Grant program, and the SFU President’s Research Start-up Grant. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WI ’17, August 23–26, 2017, Leipzig, Germany
© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.
ACM ISBN 978-1-4503-4951-2/17/08...\$15.00
<https://doi.org/10.1145/3106426.3106484>

Table 1: Example of optimal thresholds w.r.t. various tasks.

Dataset	Task	Optimal Threshold	Similarity
Wiki Editors	Spell checking	0.625	Jaccard
Restaurants	Record linkage	0.6	Jaccard
Scholar-DBLP	Record linkage	0.34	Jaccard
Wiki Links	Entity matching	0.9574	Tversky

which finds similar objects (e.g., products, people, locations) across different sources, is a powerful tool for tackling the challenge.

For example, suppose a data scientist collects a set of restaurants from Groupon.com and would like to know which restaurants are highly rated on Yelp.com. Since a restaurant may have different representations in the two data sources (e.g., “10 East Main Street” vs. “10 E Main St., #2”), she can use similarity join to find these similar restaurant pairs and integrate the two data sources together.

Threshold-driven similarity join has been extensively studied in the past [7, 9, 10, 13, 14, 18, 20–22, 26–28, 30–32]. To use it, one has to go through three steps: (a) selecting a similarity function (e.g., Jaccard), (b) selecting a threshold (e.g., 0.8), and (c) running a similarity join algorithm to find all object pairs whose similarities are at least 0.8. The existing studies are mainly focused on Step (c). However, both Steps (a) and (b) deeply implicate humans in the loop, which can be orders of magnitude slower than conducting the actual similarity join.

One may argue that, in reality, humans are able to quickly select an appropriate similarity function and a corresponding threshold for a given similarity join task. For choosing similarity function, this may be true because humans can understand the semantics of each similarity function and choose the one that meets their needs.

However, selecting an appropriate threshold may be far from easy. It is extremely difficult for humans to figure out the effect of different thresholds on result quality. Choosing a good threshold depends on not only the specified similarity function but also the underlying data. We conduct an empirical analysis on the optimal thresholds for a diverse range of similarity join tasks, where the optimal thresholds maximize F_1 -scores [29]. Table 1 shows the results (details of the experiment are in Section 5). We find that the optimal thresholds for the tasks are quite different. Even for the same similarity function, the optimal thresholds may still vary a lot. For example, the optimal threshold of a record-linkage task on the Restaurants dataset is 0.6, which differs a lot from the optimal threshold of 0.34 on the Scholar-DBLP dataset using the same similarity function.

To solve this problem, one idea may be to label some data and then use the labeled data as the ground truth of matching pairs to tune the threshold. However, human labeling is error-prone

and time-consuming, which significantly increases the (end-to-end) time of data integration or cleaning using similarity join.

In this paper, we tackle this problem from a different angle – *can we achieve high-quality similarity join results without requiring humans to label any data or specifying a similarity threshold?* Our key insight is inspired by the concept of preference in areas like economics, which is an ordering of different alternatives (results) [5]. Taking Yelp.com as an example, the restaurants can be presented in different ways such as by distance, price, or rating. The different ordering may meet different search intents. A user needs to evaluate her query intent and choose the most suitable ranking accordingly. Similarly, when performing a similarity join, we seek to provide a number of *result set preferences* for a user to select from. Intuitively, a result set preference can be thought of as an objective function to capture how much a user likes a similarity join result. Once a particular preference is chosen, we automatically tune the threshold such that the preference objective function is maximized, and then we return the corresponding similarity join result. We call this new similarity join model *preference-driven similarity join*. Compared to the traditional threshold-driven similarity join, this new model does not need any labeled data.

As a proof of concept, our paper proposes two preferences from different yet complementary perspectives. The first preference MaxGroups groups the joined pairs where each group is considered as an entity across two data sources, and returns the join result having the largest number of groups. The second preference MinOutJoin balances between matching highly similar pairs and joining many possibly matching pairs, and favors the join result minimizing the outer-join size. According to our experiments on various datasets with ground-truth, the preference-driven approach can achieve optimal or nearly optimal F_1 -scores on different tasks without knowing anything about the optimal thresholds.

Given a result set preference, a challenge is how to develop an efficient algorithm for preference-driven similarity join. This problem is more challenging than the traditional threshold-driven similarity join because it involves one additional step: finding the best threshold such that a preference is maximized. The brute-force method needs to compute the similarity values for all the pairs. It is highly inefficient even for datasets of moderate size. We solve this problem by developing a novel similarity join framework along with effective optimization techniques. The experimental results show that the proposed framework achieves several orders of magnitude speedup over the brute-force method.

The rest of the paper is organized as follows. In Section 2, we formally define the problem of preference-driven similarity join. In Section 3, we design two result set preferences from different perspectives. In Section 4, we propose a preference-driven similarity join framework, and develop efficient algorithm for set-based similarity functions. In Section 5, we evaluate our approach on four real-world web datasets from a diverse range of applications. The results suggest that preference-driven similarity join is a promising idea to tackle the threshold selection problem for similarity join, and verify that our method is effective and efficient. We review related work in Section 6 and conclude the paper in Section 7.

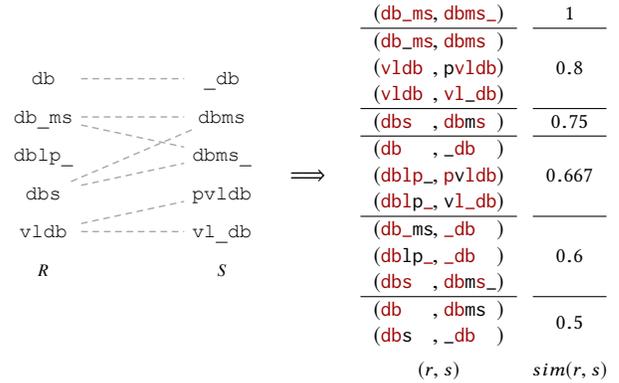


Figure 1: Example of $join(\theta)$, where $\theta = 0.5$. The optimal threshold is 0.75. The edges represent the ground-truth \mathbb{C}^+ .

2 PROBLEM DEFINITION

Let R and S be two sets of objects, \mathbb{C}^+ denote the ground-truth that is the set of pairs in $R \times S$ that should be joined/matched, and \mathbb{C}^- denote the remaining pairs, i.e., $\mathbb{C}^- = (R \times S) \setminus \mathbb{C}^+$. We call the pairs in \mathbb{C}^+ and \mathbb{C}^- matching pairs and non-matching pairs, respectively. In general, \mathbb{C}^+ and \mathbb{C}^- are assumed unknown. Figure 1 shows a toy running example of R and S , as well as the ground-truth \mathbb{C}^+ .

Let $sim : (r, s) \in R \times S \rightarrow [0, 1]$ denote a similarity function.

Definition 2.1 (Threshold-driven similarity join). Given a similarity function sim and a threshold θ , return all the object pairs (r, s) whose similarity values are at least θ , that is,

$$join(R, S, sim, \theta) = \{(r, s) \in R \times S \mid sim(r, s) \geq \theta\} \quad \square$$

If R, S , and sim are clear from the context, we write $join(\theta)$ for the sake of brevity.

We can regard $join(\theta)$ as a classifier, where the pairs returned by the function are the ones classified as positive, and the rest pairs $(R \times S) \setminus (join(\theta))$ classified as negative, that is, not-matching. Figure 1 shows an example of $join(0.5)$ using Jaccard similarity. Here, for simplicity we tokenize a string into a set of characters. For example, $jaccard(db_lp, db) = \frac{|r \cap s|}{|r \cup s|} = \frac{|{_, d, b}|}{|{_, d, b, l, p}|} = \frac{3}{5} = 0.6$.

Similarity join seeks to find a threshold that leads to the best result quality. Theoretically, there are an infinite number of thresholds to choose from. However, we only need to consider a finite subset of the possible thresholds, which is the set of the similarity values of all object pairs in $R \times S$, i.e., $\{sim(r, s) \mid r \in R, s \in S\}$, because, for any threshold not in the finite set, there is always a smaller threshold in the finite set having the same join result. For example, threshold 0.9 is not in the finite set in the example in Figure 1 but $join(0.9) = join(0.8)$.

Threshold tuning is time consuming and labor intensive. Therefore, we develop preference-driven similarity join to overcome the limitations. Before a formal definition, we first introduce the concept of *result set preference*, to capture the user preferences on a similarity join result. Formally, a result set preference is a score function $h : (R, S, sim, \theta) \rightarrow \mathbb{R}$, where \mathbb{R} is the set of real numbers. Obviously, a result set is determined by R, S, sim and θ . The result set preference gives a score on how well the result set meets a user's preference. The higher the score, the better. If R, S , and sim are clear from the context, we write $h(\theta)$ for the sake of brevity.

(db_ms, dbms_)	1	1	1
(db_ms, dbms)	0.8	2	2
(vldb, pvldb)			
(vldb, vl_db)			
(dbs, dbms)	0.75	2	2
(db, _db)			
(dbl_p_, pvldb)	0.667	3	2
(dbl_p_, vl_db)			
(db_ms, _db)			
(dbl_p_, _db)	0.6	1	-1
(dbs, dbms_)			
(db, dbms)	0.5	1	-3
(dbs, _db)			
...
$(r, s) \in R \times S$	θ	$h_c(\theta)$	$h_o(\theta)$

Figure 2: Example of h_c and h_o .

Definition 2.2 (Preference-driven Similarity Join). Given a similarity function sim and a result set preference h , return the most preferred result $join(\theta^*)$ where θ^* is the largest threshold in the finite set maximizing h . \square

For the ease of presentation, we introduce some notations. First, we denote by $join^=(\theta) = \{(r, s) \in join(\theta) \mid sim(r, s) = \theta\}$ the subset of joined pairs w.r.t. similarity θ . For example, in Figure 1, $join^=(0.75) = \{(dbs, dbms)\}$. Second, we denote by $cover^R(\theta) = \{r \mid \exists s : (r, s) \in join(\theta)\}$ the set of objects in R that are joined when the similarity threshold is θ . Similarly, we have $cover^S(\theta) = \{s \mid \exists r : (r, s) \in join(\theta)\}$. For example, in Figure 1, $cover^R(0.75) = \{db_ms, dbs, vldb\}$. Last, for $r \in R$, we denote by $top^S(r)$ the set of most similar object(s) in S including ties. Similarly, we have $top^R(s)$. For example, in Figure 1, $top^S(vldb) = \{pvldb, vl_db\}$.

3 TWO RESULT SET PREFERENCES

As a proof of concept, we present two result set preferences.

3.1 MaxGroups: Maximum Number of Non-Trivial Connected Components

Our first preference, MaxGroups, partitions objects into different groups without any prior knowledge. For a similarity threshold θ , we construct a bipartite graph $G_\theta = (U^R, V^S, join(\theta))$, where U^R and V^S are two disjoint sets of nodes representing the objects in R and S , respectively, and every pair in $join(\theta)$ defines an edge. As indicated in [23], each connected component in the bipartite graph is an entity, where the objects in the same connected component are the entity’s different representations.

MaxGroups prefers the similarity join result with more non-trivial connected components (i.e., connected components with at least two nodes, one in R and another in S). The intuition is that heuristically we want to match as many entities as possible across R and S . Let $J(G_\theta)$ denote the set of non-trivial connected components in a bipartite G_θ , we define result set preference MaxGroups as

$$h_c(\theta) = |J(G_\theta)|$$

Figure 2 gives an example of h_c on our toy dataset in Figure 1.

Given a similarity join result $join(\theta)$, the time complexity of computing $h_c(\theta)$ is $O(|R| + |S| + |join(\theta)|)$ by simply computing the

$$\begin{aligned}
 & (db_ms, dbms) \\
 & (db_ms, dbms) \\
 & (dbs, dbms) \cup (db, \text{NULL}) \cup (\text{NULL}, _db) \\
 & (vldb, pvldb) \cup (dbl_p_, \text{NULL}) \cup (\text{NULL}, _db) \\
 & (vldb, vl_db)
 \end{aligned}$$

Figure 3: Example of $outjoin(\theta)$, where $\theta = 0.75$.

connected components of the bipartite graph G_θ . Since we need to compute the preference scores for multiple similarity thresholds θ , there are opportunities to reduce the computational cost. We will discuss the details in Section 4.3.

3.2 MinOutJoin: Minimum Outer-Join Size

We make the following observation. On the one hand, if we set a too high similarity threshold and thus be too strict in similarity, many objects may not be matched with their counterparts due to noise. The extreme case is that, if we set the similarity threshold to 1, only those perfectly matching objects are joined. On the other hand, if we set a too low similarity threshold and thus be too loose in similarity, many not-matching objects may be joined by mistake. The extreme case is that, by setting the similarity threshold to 0, every pair of objects in the two sets are joined.

We need to find a good balance between the two and strive to a good tradeoff. Technically, full outer-join includes both joined entries and those not joined (by matching with NULL). The size of the full outer-join is jointly determined by the number of objects joined and the number of objects not joined. The two numbers trade off each other. Therefore, if we minimize the size of the full outer-join, we reach a tradeoff between the two ends. This is the intuition behind our second preference, MinOutJoin.

The full outer similarity join result w.r.t. a similarity threshold is

$$\begin{aligned}
 outjoin(\theta) = & join(\theta) \cup \left\{ (r, \text{NULL}) \mid r \in R \setminus cover^R(\theta) \right\} \\
 & \cup \left\{ (\text{NULL}, s) \mid s \in S \setminus cover^S(\theta) \right\}
 \end{aligned}$$

where $\{(r, \text{NULL}) \mid r \in R \setminus cover^R(\theta)\}$ is the set of objects in R that are not joined, and $\{(\text{NULL}, s) \mid s \in S \setminus cover^S(\theta)\}$ is the set of objects in S that are not joined. Figure 3 illustrates an example of a full outer-join. We define our preference MinOutJoin as

$$h_o(\theta) = |R| + |S| - |outjoin(\theta)| = |cover^R(\theta)| + |cover^S(\theta)| - |join(\theta)|$$

where $|R| + |S|$ is a constant given R and S . Figure 2 gives an example of h_o on our toy dataset.

This preference gives a penalty when multiple objects in a set are joined with multiple objects in the other set. Joining x objects in R and y objects in S results in $x \cdot y$ pairs in the full outer-join. Not joining them results in at most $x + y$ of pairs in the full outer-join. When $x > 1$ and $y > 1$, we have $x \cdot y \geq x + y$.

Given a threshold θ , it is straightforward to compute $cover^R(\theta)$, $cover^S(\theta)$ and $join(\theta)$ according to their definitions by scanning the join result $join(\theta)$. The time complexity of computing $h_o(\theta)$ is $O(|join(\theta)|)$ for each θ . Since we need to compute the preference scores for multiple similarity thresholds θ , there are opportunities to reduce the computational cost. We will discuss the details in Section 4.3.

Algorithm 1: Preference-driven similarity join framework.

Input: objects R and S , similarity function sim , preference h

Output: the most preferred join result $join(\theta^*)$

```
1  $\Theta \leftarrow \text{PivotalThresholds}(R, S, sim)$ 
2 foreach threshold  $\theta_i \in \Theta$  in descending order do
3    $join^=(\theta_i) \leftarrow \text{IncrementalSimJoin}(\theta_{i-1}, \theta_i)$ 
4    $join(\theta_i) \leftarrow join(\theta_{i-1}) \cup join^=(\theta_i)$ 
5   if  $h(\theta_i) > h(\theta^*)$  then  $\theta^* \leftarrow \theta_i$  // INCREMENTALSCORE
6   else if  $\widehat{h}(\theta_i) \leq h(\theta^*)$  then break // EARLYTERMINATION
7 return  $join(\theta^*)$ 
```

4 ALGORITHM FRAMEWORK

In this section, we present an efficient framework for the preference-driven similarity join problem. A brute-force solution is to compute the similarities for all the object pairs, calculate the preference score w.r.t. each possible threshold, and return the similarity join result with the highest preference score. This brute-force method may be inefficient. Computing similarities for all pairs is often prohibitive. The number of possible thresholds can be very large, $|R| \times |S|$ in the worst case. It is crucial to reduce the cost involved in this process.

To tackle the challenges, we propose a preference-driven similarity join framework in Algorithm 1. Central to the framework are four key functions. Function `PivotalThresholds` (Section 4.1) identifies a small set of thresholds Θ , called the pivotal thresholds, that are guaranteed to cover the best preference score obtained from all the possible thresholds. Function `IncrementalSimJoin` (Section 4.2) checks the pivotal thresholds in value descending order and incrementally computes the similarity join result for each threshold. We propose a new optimization technique, called lazy evaluation, to further improve the efficiency. Function `IncrementalScore` (Section 4.3) computes the preference score for each threshold. It is possible to reduce the cost by computing the scores incrementally when checking the pivotal thresholds in value descending order. Function `EarlyTermination` (Section 4.4) determines whether we can stop checking the remaining pivotal thresholds by comparing the upper bound $\widehat{h}(\theta_i)$ with currently best score $h(\theta^*)$ once a similarity join result $join(\theta_i)$ is computed.

4.1 Pivotal Thresholds

Not every threshold has a chance to lead to the maximum preference score. In this section, we study how to identify a small set of thresholds Θ such that the maximum preference score can be obtained by only evaluating Θ , i.e., $\max_{\theta \in [0,1]} h(\theta) = \max_{\theta \in \Theta} h(\theta)$.

Consider $\Theta = \{sim(r, s) \mid r \in R, s \in S, r \in top^R(s) \wedge s \in top^S(r)\}$. Clearly, Θ is often dramatically smaller than $|R| \times |S|$, as

$$|\Theta| \leq \min \left\{ \left| \left\{ sim(r, s) \mid r \in R, s \in S \text{ where } s \in top^S(r) \right\} \right|, \left| \left\{ sim(r, s) \mid r \in R, s \in S \text{ where } r \in top^R(s) \right\} \right| \right\} \leq \min \{|R|, |S|\}$$

For example, in Figure 1, $\Theta = \{1, 0.8, 0.667\}$.

We can show that both `MaxGroups` and `MinOutJoin` have the same set of pivotal thresholds. The basic idea is that, for any $\theta \notin \Theta$, there exists $\theta' > \theta$ such that $h(\theta') \geq h(\theta)$. Remind that, through the paper, we only need to discuss the thresholds within the finite set $\{sim(r, s) \mid r \in R, s \in S\}$ as discussed in Section 2.

LEMMA 4.1. *Given a threshold θ , if $r \notin top^R(s)$ or $s \notin top^S(r)$ for any $(r, s) \in join^=(\theta)$, then $\exists \theta' > \theta : h_c(\theta') \geq h_c(\theta)$.*

PROOF. Let θ' be a threshold such that $join(\theta) \setminus join(\theta') = join^=(\theta)$. Bipartite G_θ can be derived by adding those new edges in $join^=(\theta)$ to bipartite $G_{\theta'}$. For any $(r, s) \in join^=(\theta)$, if $r \notin top^R(s)$, then r must be already in a non-trivial connected component of $G_{\theta'}$, and s can only be added to the non-trivial connected component where r belongs to. Similar situation happens if $s \notin top^S(r)$. Since there is not any new non-trivial connected component in G_θ comparing to $G_{\theta'}$, $h_c(\theta') \geq h_c(\theta)$. \square

LEMMA 4.2. *Given a threshold θ , if $r \notin top^R(s)$ and $s \notin top^S(r)$ for any $(r, s) \in join^=(\theta)$, then $\exists \theta' > \theta : h_o(\theta') \geq h_o(\theta)$.*

PROOF. Let θ' be a threshold such that $join(\theta) \setminus join(\theta') = join^=(\theta)$. When $r \notin top^R(s) \vee s \notin top^S(r)$ for any $(r, s) \in join^=(\theta)$,

$$\begin{aligned} & |join(\theta)| - |join(\theta')| = |join^=(\theta)| \\ & \geq \left| \left\{ (r, s) \in join^=(\theta) \mid s \in top^S(r) \right\} \right| + \left| \left\{ (r, s) \in join^=(\theta) \mid r \in top^R(s) \right\} \right| \\ & \geq \left| \left\{ r \in cover^R(\theta) \mid s \in S \text{ where } sim(r, s) = \theta \wedge s \in top^S(r) \right\} \right| \\ & \quad + \left| \left\{ s \in cover^S(\theta) \mid r \in R \text{ where } sim(r, s) = \theta \wedge r \in top^R(s) \right\} \right| \\ & = |cover^R(\theta)| - |cover^R(\theta')| + |cover^S(\theta)| - |cover^S(\theta')| \end{aligned}$$

Thus, $h_o(\theta') - h_o(\theta) \geq 0$. \square

Thanks to the existing fast top- k similarity search algorithms [31, 33], obtaining the pivotal thresholds can be efficient by computing the most similar objects of each object in R and S , respectively. According to the above lemmas, it is guaranteed that the largest threshold having the maximum score in the finite set of thresholds is always in Θ .

4.2 Incremental Similarity Join

In this section, we present an efficient algorithm that incrementally computes the similarity join result $join(\theta_i)$ w.r.t. threshold θ_i . We do not need to conduct a similarity join for each pivotal threshold. Since $join(\theta) = \cup_{\theta' \geq \theta} join^=(\theta')$, for each threshold θ , we only need to compute the respective newly joined pairs $join^=(\theta)$. Thus, we can enumerate the thresholds in the value descending order, and compute the respective join results incrementally.

The algorithm consists of two steps. First, the algorithm incrementally computes a set of candidate pairs $cand(\theta_i)$. Second, the algorithm evaluates the similarity of each candidate pair and returns the pairs whose similarity values are at least the threshold θ_i . While this two-step approach has been used by existing similarity join algorithms [9, 13, 28, 32], our contribution is a novel optimization technique, called lazy evaluation, which lazily evaluate the similarity of each candidate pair and reduces the cost.

We focus on set-based similarity functions in this paper. Similar strategies can be applied to other kinds of similarity functions, like string-based or vector-based. Note that multi-set (bag) can also be used here instead of set. Table 2 shows the definitions of the similarity functions. Jaccard, overlap, dice, and cosine similarity are widely adopted in existing similarity join literature [9, 13, 28, 32]. In addition, we include Tversky similarity [25], which is a special

Table 2: Summary of set-based similarity functions. Here $bound_{i,j}^{min} = |r[:i] \cap s[:j]| \leq |r \cap s|$ and $bound_{i,j}^{max} = |r[:i] \cap s[:j]| + \min\{|r| - i, |s| - j\} \geq |r \cap s|$.

Similarity	sim	$sim_{i,j}^{min}$	$sim_{i,j}^{max}$	t_θ
Jaccard	$\frac{ r \cap s }{ r + s - r \cap s }$	$\frac{bound_{i,j}^{min}}{ r + s - bound_{i,j}^{min}}$	$\frac{bound_{i,j}^{max}}{ r + s - bound_{i,j}^{max}}$	$[\theta \cdot r]$
Overlap	$\frac{ r \cap s }{\max\{ r , s \}}$	$\frac{bound_{i,j}^{min}}{\max\{ r , s \}}$	$\frac{bound_{i,j}^{max}}{\max\{ r , s \}}$	$[\theta \cdot r]$
Dice	$\frac{2 \cdot r \cap s }{ r + s }$	$\frac{2 \cdot bound_{i,j}^{min}}{ r + s }$	$\frac{2 \cdot bound_{i,j}^{max}}{ r + s }$	$[\frac{\theta}{2} \cdot r]$
Cosine	$\frac{ r \cap s }{\sqrt{ r \cdot s }}$	$\frac{bound_{i,j}^{min}}{\sqrt{ r \cdot s }}$	$\frac{bound_{i,j}^{max}}{\sqrt{ r \cdot s }}$	$[\theta^2 \cdot r]$
Tversky	$\frac{ r \cap s }{\alpha \cdot r + (1 - \alpha) \cdot s }$	$\frac{bound_{i,j}^{min}}{\alpha \cdot r + (1 - \alpha) \cdot s }$	$\frac{bound_{i,j}^{max}}{\alpha \cdot r + (1 - \alpha) \cdot s }$	$[\theta \cdot \alpha \cdot r]$

asymmetric set-based similarity with different weights α and $1 - \alpha$ on r and s , respectively. This similarity function is very useful in certain scenarios, like matching a text with an entity contained by the text. Given an object r as a set, we use $r[:i]$ to denote the first i elements of r assuming a global ordering of elements in the set.

4.2.1 Candidate Pair Generation. Established by prefix filtering [13], if $sim(r, s) \geq \theta$, the number of elements in the overlap of the sets $|r \cap s|$ is no fewer than an overlap threshold t_θ w.r.t. $|r|$, where the overlap thresholds for set-based similarities are shown in Table 2. Thus, the candidate pair generation problem is converted to how to filter out the pairs with fewer than t_θ common elements.

To filter out the pairs with less than t_θ common elements, we fix a global ordering on the elements of all the objects, and sort the elements in each object based on the ordering. Like [28], we use the inverse document frequency as the global ordering. Prefix filtering [13] establishes that if $|r \cap s| \geq t_\theta$, then $r[:\#prefix_\theta(r)] \cap s[:\#prefix_\theta(r)] \neq \emptyset$, where $\#prefix_\theta(r) = |r| - t_\theta + 1$.

Using an inverted index, we do not need to enumerate each pair (r, s) to verify whether $r[:\#prefix_\theta(r)] \cap s[:\#prefix_\theta(r)] \neq \emptyset$. An inverted index maps an element to a list of objects containing the element. After building the inverted index for S , for each $r \in R$, we only need to merge the inverted lists of the elements in $r[:\#prefix_\theta(r)]$ to retrieve each $s \in S$ such that $r[:\#prefix_\theta(r)] \cap s[:\#prefix_\theta(r)] \neq \emptyset$.

Our goal is to generate the candidate pairs for $[\theta_i, \theta_{i-1}]$. We use an incremental prefix filtering approach [31] that memorizes previous results to avoid regenerating the candidate pairs for $[\theta_{i-1}, 1]$.

4.2.2 Lazy Evaluation. Suppose we want to check whether the similarity of a candidate pair $(r, s) \in cand(\theta_i)$ is no smaller than a threshold θ_i or not. The idea of lazy evaluation is to iteratively compute both a maximum and a minimum possible value of $sim(r, s)$, denoted by $sim^{max}(r, s)$ and $sim^{min}(r, s)$, respectively. Interestingly, both $sim^{max}(r, s)$ and $sim^{min}(r, s)$ get tighter through the process, and finally converge at $sim(r, s)$. During this process, we use the values for lazy evaluation in two ways.

- If $sim^{max}(r, s) < \theta_i$, then $sim(r, s) < \theta_i$. Thus, it is only necessary to resume evaluating (r, s) for a future smaller threshold θ_j where $\theta_{j-1} > sim^{max}(r, s) \geq \theta_j$.
- If $sim^{min}(r, s) \geq \theta_i$, then $\theta_{i-1} > sim(r, s) \geq \theta_i$. Thus, (r, s) does not need to be fully evaluated at all.

We scan r and s iteratively together from left to right, according to the global ordering. Assuming $r[:i]$ and $s[:j]$ have been scanned, Table 2 shows the maximum/minimum possible values of

Algorithm 2: Incremental similarity join.

Input: thresholds θ_{i-1} and θ_i where $\theta_{i-1} > \theta_i$
Output: incremental similarity join result $join^=(\theta_i)$

- 1 $join^=(\theta_i) \leftarrow \emptyset$
- 2 let $cand(\theta_i)$ be the candidate pairs for $[\theta_i, \theta_{i-1}]$
- 3 **foreach** pair $(r, s) \in cand(\theta_i)$ **do**
- 4 **while** $sim^{min}(r, s) < \theta_i \leq sim^{max}(r, s)$ **do**
- 5 update $sim^{max}(r, s)$ and $sim^{min}(r, s)$
- 6 **if** $sim^{max}(r, s) < \theta_i$ **then**
- 7 find $\theta_j : \theta_{j-1} > sim^{max}(r, s) \geq \theta_j$ by binary search
- 8 add (r, s) into $cand(\theta_j)$
- 9 **else**
- 10 add (r, s) into $join^=(\theta_i)$
- 11 **return** $join^=(\theta_i)$

$sim(r, s)$. Through the scanning, we iteratively update $sim^{max}(r, s)$ and $sim^{min}(r, s)$ accordingly.

The pseudo-code of the lazy evaluation-powered algorithm is shown in Algorithm 2. The algorithm first computes $cand(\theta_i)$, by generating a set of candidate pairs for $[\theta_i, \theta_{i-1}]$ together with the previously postponed candidate pairs from larger thresholds. Then, the algorithm examines each candidate pair in $cand(\theta_i)$ and decides whether it should be added into $join^=(\theta_i)$ or postponed and added into $cand(\theta_j)$ for a smaller threshold θ_j (found by binary search over the rest of the thresholds). Finally, $join^=(\theta_i)$ is returned.

4.3 Incremental Score Computation

For both our preferences, computing the preference score for each similarity threshold is straightforward. However, as we need to compute the preference scores for multiple thresholds, it is necessary to explore how to further reduce the cost.

For preference MaxGroups, when computing the join result incrementally by decreasing θ , if two objects of each newly joined pair in $join^=(\theta)$ are in different connected components, the connected components are merged together to form a larger connected component. We use a disjoint-set data structure to dynamically track newly joined pairs, and update the connected components accordingly. It only takes almost $O(1)$ amortized time [15] for each newly joined pair in $join^=(\theta)$.

For preference MinOutJoin, when processing incrementally by the value decreasing order of θ , we only need to scan each $join^=(\theta)$ to update $join(\theta)$, $cover^R(\theta)$, $cover^S(\theta)$, and further the preference score. It only takes $O(1)$ time for each newly joined pair in $join^=(\theta)$.

4.4 Early Termination

The goal of early termination is to determine if we can return the current most preferred result without evaluating the remaining thresholds. In our algorithm, the thresholds are evaluated in the descending order. Suppose threshold θ_i has just been evaluated. At this point, we have known the preference score for each threshold that is at least θ_i . Let $h(\theta^*)$ denote the current best preference score, i.e., $h(\theta^*) = \max_{\theta \geq \theta_i} h(\theta)$. If we can derive an upper-bound $\hat{h}(\theta_i)$ of the preference scores for the remaining thresholds and show that the upper-bound is no larger than $h(\theta^*)$, then it is safe to stop at θ_i and $h(\theta^*)$ is the best result overall.

Table 3: Dataset characteristics.

Dataset	R				S				$ \mathbb{C}^+ $
	$ R $	Max. Len.	Avg. Len.	$ S $	Max. Len.	Avg. Len.	$ S $		
Wiki Editors	2,239	20	9.65	1,922	16	8.66	2,455		
Restaurants	533	96	48.38	331	91	43.5	112		
Scholar-DBLP	64,259	259	115.9	2,562	326	106.61	5,347		
Wiki Links	187,122	1,393	17.08	168,652	209	16.35	202,272		

For MaxGroups, as the threshold decreases, a previously unseen non-trivial connected component can only be created by merging two trivial connected components. Since a new non-trivial connected component contains at least one object from $R \setminus \text{cover}^R(\theta_i)$ and one object from $S \setminus \text{cover}^S(\theta_i)$, the number of non-trivial connected components in any $G_{\theta'}$ such that $\theta' < \theta_i$ is at most $\min\{|R \setminus \text{cover}^R(\theta_i)|, |S \setminus \text{cover}^S(\theta_i)|\}$. Thus, an upper-bound is

$$\widehat{h}_c(\theta_i) = h_c(\theta_i) + \min\{|R \setminus \text{cover}^R(\theta_i)|, |S \setminus \text{cover}^S(\theta_i)|\}$$

For MinOutJoin, as the threshold decreases, the join result $\text{join}(\theta_i)$ includes more pairs. Whenever a new pair (r, s) is joined, $|\text{join}(\theta_i)|$ increases by one. The only way to get the preference score increased by 1 is that $|\text{cover}^R(\theta_i)|$ and $|\text{cover}^S(\theta_i)|$ both increase by 1. In this case, r has to be come from $R \setminus \text{cover}^R(\theta_i)$ and s has to be come from $S \setminus \text{cover}^S(\theta_i)$. Therefore, the preference score can at most increase by $\min\{|R \setminus \text{cover}^R(\theta_i)|, |S \setminus \text{cover}^S(\theta_i)|\}$. Accordingly, we set an upper-bound to

$$\widehat{h}_o(\theta_i) = h_o(\theta_i) + \min\{|R \setminus \text{cover}^R(\theta_i)|, |S \setminus \text{cover}^S(\theta_i)|\}$$

5 EXPERIMENTAL RESULTS

We present a series of experimental results in this section. The programs were implemented in Python running with PyPy¹. The experiments were conducted using a Mac Pro Late 2013 Server with Intel Xeon 3.70GHz CPU, 64GB memory, and 256GB SSD.

5.1 Datasets

We adopt four real-world web datasets with ground-truth for evaluation. Table 3 shows the characteristics of each dataset.

Wiki Editors [1] is about misspellings of English words, made by Wikipedia page editors where the errors are mostly typos. Each misspelling has at least one correct word. R contains the misspellings, while S contains the correct words. The ground-truth \mathbb{C}^+ are pairs of each misspelling and the corresponding correct word.

Restaurants [2] links the restaurant profiles between two websites. Each profile contains the name and address of a restaurant. We remove the phone number and cuisine type, which are available in the original data, to make it more challenging. R and S are profiles of restaurants, and the ground-truth \mathbb{C}^+ identifies the pairs of profiles linking the same restaurants. Every restaurant has at most one match.

Scholar-DBLP [3] finds the same publications in Google Scholar and DBLP, where each record in DBLP has at least one matching record in Google Scholar. Each record on both websites contains the title, author names, venue, and year. R and S are publications identified by Google Scholar and DBLP, respectively, and the ground-truth \mathbb{C}^+ are pairs of records linking the same publications.

¹PyPy (<http://pypy.org/>) is an advanced just-in-time compiler, providing 10 times faster performance for our algorithm than the standard Python interpreter.

Table 4: Sensitivity of thresholds on F_1 -score. Three thresholds are used: the optimal one $\theta^\#$, the higher one $\theta^+ = \min\{1, \theta^\# + 0.1\}$, and the lower one $\theta^- = \max\{0, \theta^\# - 0.1\}$. Numbers in red are when ours achieves optimal F_1 -scores.

Dataset	using θ^+	using $\theta^\#$	using θ^-	Preference-driven
Wiki Editors	0.599	0.764	0.705	0.764 using h_c and h_o
Restaurants	0.725	0.816	0.597	0.809 using h_o
Scholar-DBLP	0.697	0.759	0.554	0.754 using h_c
Wiki Links	0.774	0.780	0.624	0.780 using h_c

Wiki Links [4] is a large dataset containing short anchor text on web pages and the Wikipedia link that each anchor text contains. R contains the anchor text, while S contains the Wikipedia entities extracted from Wikipedia links. For example, link [https://en.wikipedia.org/wiki/Harry_Potter_\(character\)](https://en.wikipedia.org/wiki/Harry_Potter_(character)) is converted to entity “Harry Potter”. The ground-truth \mathbb{C}^+ are pairs linking each anchor text and its entity. Each anchor text may also contain the text of the entity that the Wikipedia link refers to, such as “...Harry Potter is a title character...”. This dataset has multiple files, where each contains roughly 200,000 mappings (about 200 MB). We used the first file in most of our evaluation except for the scalability evaluation where the first four files were used.

To adopt set-based similarities, each string needs to be converted to a set. Empirically, people often convert a long string into a bag of words and convert a short string into a set of grams. For the first dataset, since a string represents a word, we convert each string into a bag of 2-grams. For the rest three datasets, since the strings are much longer, we convert each string into a bag of words. We use Jaccard similarity for the first three datasets. For the Wiki Links dataset, since anchor text often contains both the entity text and some other irrelevant texts, we choose Tversky similarity ($\alpha = 0.1$).

5.2 Threshold-driven vs. Preference-driven

In this section, we empirically investigate the pros and cons of both our preference-driven approach and other possible solutions for tuning a similarity join threshold.

We demonstrate the sensitivity of thresholds w.r.t. F_1 -score in Table 4. A small deviation from the optimal threshold affects the result quality dramatically. This clearly shows that threshold tuning is crucial for threshold-driven similarity join. On two datasets, our preference-driven approach can achieve the optimal F_1 -score, and on the other two datasets, they are very close to the optimal.

To compare with human labeling approaches, we adopt two supervised approaches to tune a similarity threshold. The first one is random sampling, where pairs are randomly sampled and labeled. We use the same sampling method as [17], where the larger set R is sampled with a specified sampling rate, and then joined with S to derive all the pairs to be labeled. Apparently, at least $|S|$ pairs need to be labeled. The second approach [11] uses active learning to tune a threshold by incrementally labeling the most uncertain pair to the classifier. When there is a tie, the one with the highest similarity is selected. The threshold that achieves the best on the labeled data is selected. Figure 4 shows the results. For random sampling, only using a very high sampling rate like 10% can almost catch up with our method. For active learning, the number of necessary labels is significantly reduced, however, still hundreds to thousands of pairs need to be labeled on most of the datasets.

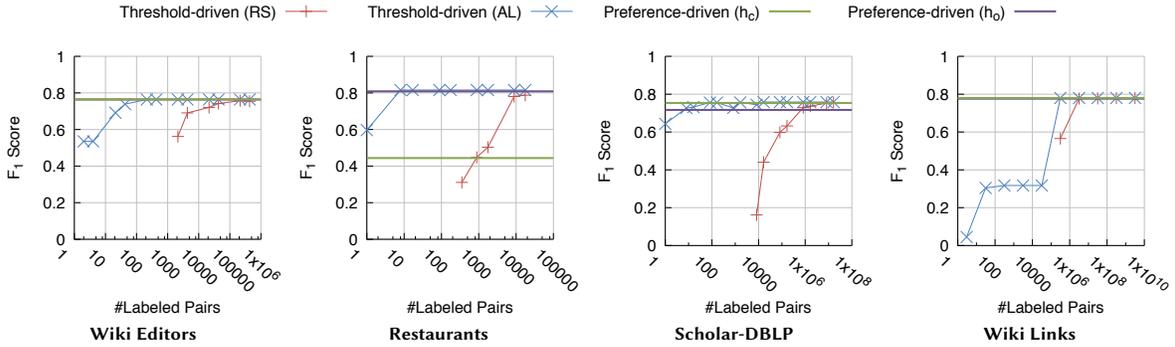


Figure 4: Comparisons of accuracy between threshold-driven (using random sampling (RS) or active learning (AL) for threshold tuning) and preference-driven approaches.

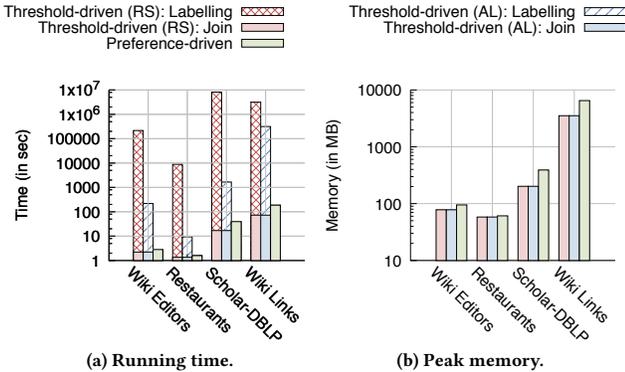


Figure 5: Comparisons of efficiency between threshold-driven and preference-driven approaches. For threshold-driven approach, we tune the threshold such that it achieves the closest (within 0.01) F_1 -score as preference-driven approach. We assume that each pair needs 1s to label. The same preference in Table 4 is used here for each dataset.

The total running time of the two supervised approaches contains two parts: the labeling part which tunes the threshold, and the joining part using the tuned threshold. In comparison, our approach returns both a threshold and its join results in a unified framework. We calculate the end-to-end time that the threshold-driven approach needs in order to achieve the same quality as our preference-driven approach. For simplicity, we assume that each pair takes 1s to be labeled correctly by a human, which is a very conservative estimation.

Figure 5 shows the results. The labeling step is much more costly than the joining step. The end-to-end time of the preference-driven approach is orders of magnitude faster. The preference-driven approach uses one to two times more memory than a threshold-driven approach, due to the need of caching the candidate pairs.

5.3 Accuracy of Preference-driven Approach

In this section, we evaluate the accuracy of our preference-driven approach. Table 5 compares the results between h_c and h_o . The precision, recall, and F_1 -score are presented, together with the preferred threshold θ^* . For Wiki Editors, due to its nature of high similarity between misspellings and correct words, both preferences return the same optimal result. For the record-linkage task

Table 5: Accuracy for varying datasets and preferences.

Dataset	h_c (MaxGroups)			h_o (MinOutJoin)				
	θ^*	Precision	Recall F_1	θ^*	Precision	Recall F_1		
Wiki Editors	0.625	0.837	0.704	0.764	0.625	0.837	0.704	0.764
Restaurants	0.429	0.291	0.938	0.444	0.556	0.805	0.813	0.809
Scholar-DBLP	0.361	0.841	0.683	0.754	0.419	0.903	0.595	0.717
Wiki Links	0.957	0.959	0.657	0.780	0.972	0.968	0.648	0.776

on Restaurants, h_o gives the significantly better result as it favors one-to-one matching. For Scholar-DBLP, h_c gives the better result, because maximizing the number of non-trivial connected components actually satisfies the nature of many-to-many matching. For the Wiki Links, h_c gives the optimal result, and h_o is quite close.

5.4 Efficiency of Preference-driven Approach

There is no existing work solving the same problem. We choose the brute-force method in Section 4 as a baseline to evaluate the efficiency. This method takes almost the same time and memory regardless of the preference due to the same amortized time for processing each newly joined pair incrementally.

Figure 6(a) shows the number of thresholds evaluated by the baseline and our algorithm. On most of the datasets, our algorithm evaluates 10 to 100 times less thresholds than the baseline. Our method achieves a significant speedup due to the combination of other optimization techniques (i.e., incremental similarity join and early termination). As Figure 6(b) shows, our algorithm is 10 to 100 times faster than the baseline on all the large datasets. Figure 6(c) shows that our method consumes significantly less memory.

For incremental similarity join, we use lazy evaluation for speedup. Instead, a simple approach computes the exact similarities for all the new candidate pairs produced by each threshold, and puts them into a max-heap. Those pairs in the heap whose similarities are no less than the current threshold are popped out for evaluation. On smaller datasets Wiki Editors and Restaurants, the simple approach works slightly better, as there are not many pairs that need lazy evaluation. However, on larger datasets Scholar-DBLP and Wiki Links, our lazy evaluation approach is 2 to 5 times faster.

For scalability, we evaluate our algorithm using h_c on a larger version of Wiki Links that contains 4 files. There are overlapping objects between different parts. Figure 7 shows the scalability results on number of evaluated thresholds, running time, and peak memory usage. Our method is scalable.

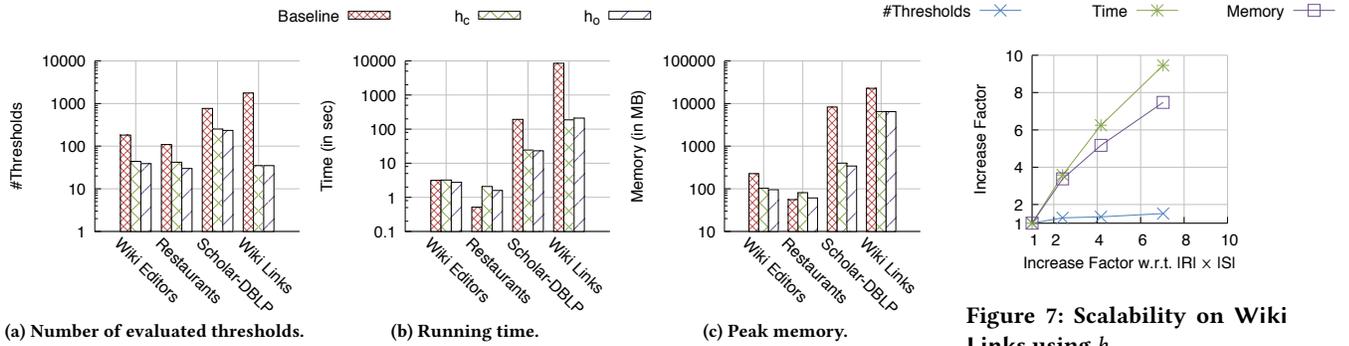


Figure 6: Efficiency of preference-driven approach.

6 RELATED WORK

Due to the crucial role of similarity join in data integration and data cleaning, a large number of similarity join algorithms have been proposed [7, 9, 13, 18, 20, 27, 28, 30–32]. There are also scalable implementations of the algorithms using the MapReduce framework [14, 22, 26]. Top-k similarity join is also explored [19, 31, 33].

While the majority of the existing work on similarity join needs to specify a similarity threshold or a limit of the number of results returned, there do exist some studies that seek to find a suitable threshold for similarity join in a supervised way [8, 11, 12, 29]. Both [8] and [11] adopted active learning to tweak the threshold. Chaudhuri *et al.* [12] learned an operator tree, where each node contains a similarity threshold and a similarity function on each of the splitting attributes. Wang *et al.* [29] modeled this problem as an optimization problem and applied hill climbing to optimize the threshold-selection process. Nevertheless, all these methods need humans to label a number of pairs, which are selected based on either random sampling or active learning. In comparison, our method does not need any labeled data.

There are studies about preferences in the database research (See [24] for a thorough survey.). However, they mainly focus on how the joined tuples are ranked and selected, instead of how the tables (and in our case sets of objects) are joined to generate the joined tuples. Alexe *et al.* [6] discussed user-defined preference rules for integrating temporal data, which is orthogonal to our work. To the best of our knowledge, [16] is the only work discussing result set preference for joining relational tables.

7 CONCLUSIONS

Threshold selection, usually neglected in the similarity join literature, can actually be a bottleneck in an end-to-end similarity join process. To mitigate the challenge, we propose preference-driven similarity join in this paper. We formalize preference-driven similarity join, and propose the notion of result set preference. We present two specific preferences, MaxGroups and MinOutJoin as proofs of concept. We develop a general framework for preference-driven similarity join and the efficient computation methods. We evaluate our approach on four real-world datasets from a diverse range of application scenarios. The results demonstrate that preference-driven similarity join can achieve high-quality results without any labeled data, and our proposed framework is efficient and scalable. Future directions include developing more preferences for various application scenarios and supporting various similarity functions.

REFERENCES

- [1] <http://www.dcs.bbk.ac.uk/~ROGER/corpora.html>.
- [2] <http://www.cs.utexas.edu/users/ml/riddle/data.html>.
- [3] http://dbs.uni-leipzig.de/en/research/projects/object_matching.
- [4] <http://www.iesl.cs.umass.edu/data/wiki-links>.
- [5] K. J. Arrow, M. D. Intriligator, W. Hildenbrand, and H. Sonnenschein. Handbook of mathematical economics. *Handbook of mathematical economics*, 1991.
- [6] B. Alexe, M. Roth, and W. Tan. Preference-aware Integration of Temporal Data. *PVLDB*, 8(4):365–376, 2014.
- [7] A. Arasu, V. Ganti, and R. Kaushik. Efficient exact set-similarity joins. In *VLDB’2006*.
- [8] A. Arasu, M. Götz, and R. Kaushik. On active learning of record matching packages. In *SIGMOD’2010*.
- [9] R. J. Bayardo, Y. Ma, and R. Srikant. Scaling up all pairs similarity search. In *WWW’2007*.
- [10] P. Bouros, S. Ge, and N. Mamoulis. Spatio-textual similarity joins. *PVLDB*, 6(1):1–12, 2012.
- [11] L. Büch and A. Andrzejak. Approximate string matching by end-users using active learning. In *CIKM’2015*.
- [12] S. Chaudhuri, B. Chen, V. Ganti, and R. Kaushik. Example-driven design of efficient record matching queries. In *VLDB’2007*.
- [13] S. Chaudhuri, V. Ganti, and R. Kaushik. A primitive operator for similarity joins in data cleaning. In *ICDE’2006*.
- [14] D. Deng, G. Li, S. Hao, J. Wang, and J. Feng. Massjoin: A mapreduce-based method for scalable string similarity joins. In *ICDE’2014*.
- [15] D. Eppstein, Z. Galil, and G. F. Italiano. Dynamic graph algorithms. 1998.
- [16] C. Gao, J. Pei, J. Wang, and Y. Chang. Schemaless Join for Result Set Preferences. In *IRI’2017*.
- [17] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. W. Shavlik, and X. Zhu. Corlone: hands-off crowdsourcing for entity matching. In *SIGMOD’2014*.
- [18] Y. Jiang, G. Li, J. Feng, and W. Li. String similarity joins: An experimental evaluation. *PVLDB*, 7(8):625–636, 2014.
- [19] Y. Kim and K. Shim. Parallel top-k similarity join algorithms using mapreduce. In *ICDE’2012*.
- [20] G. Li, D. Deng, J. Wang, and J. Feng. Pass-join: A partition-based method for similarity joins. *PVLDB*, 5(3):253–264, 2011.
- [21] J. Lu, C. Lin, W. Wang, C. Li, and H. Wang. String similarity measures and joins with synonyms. In *SIGMOD’2013*.
- [22] A. Metwally and C. Faloutsos. V-smart-join: A scalable mapreduce framework for all-pair similarity joins of multisets and vectors. *PVLDB*, 5(8):704–715, 2012.
- [23] A. E. Monge and C. Elkan. An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. In *DMKD’1997*.
- [24] K. Stefanidis, G. Koutrika, and E. Pitoura. A survey on representation, composition and application of preferences in database systems. *ACM TODS*, 36(3):19, 2011.
- [25] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.
- [26] R. Vernica, M. J. Carey, and C. Li. Efficient parallel set-similarity joins using mapreduce. In *SIGMOD’2010*.
- [27] J. Wang, G. Li, and J. Feng. Trie-join: Efficient trie-based string similarity joins with edit-distance constraints. *PVLDB*, 3(1):1219–1230, 2010.
- [28] J. Wang, G. Li, and J. Feng. Can we beat the prefix filtering?: an adaptive framework for similarity join and search. In *SIGMOD’2012*.
- [29] J. Wang, G. Li, J. X. Yu, and J. Feng. Entity matching: How similar is similar. *PVLDB*, 4(10):622–633, 2011.
- [30] C. Xiao, W. Wang, and X. Lin. Ed-join: an efficient algorithm for similarity joins with edit distance constraints. *PVLDB*, 1(1):933–944, 2008.
- [31] C. Xiao, W. Wang, X. Lin, and H. Shang. Top-k set similarity joins. In *ICDE’2009*.
- [32] C. Xiao, W. Wang, X. Lin, and J. X. Yu. Efficient similarity joins for near duplicate detection. In *WWW’2008*.
- [33] X. Xu, C. Gao, J. Pei, K. Wang, and A. Al-Barakati. Continuous similarity search for evolving queries. *KAIS*, 48(3):649–678, 2016.

Figure 7: Scalability on Wiki Links using h_c .