1 Minwise and Maxwise Hashing

Minwise hashing, as described in [?], has repeatedly proven a powerful tool when comparing large sets of strings rapidly, especially for duplicate detection of long articles. The use of minwise hashing for rRNA sequences has already been done in [?], however the method of this paper will be extended by applying two methods of maxwise hashing as described in [?].

1.1 Introduction to Minwise Hashing

Consider two sets A and B. To find the similarity between the two sets, minwise hashing uses the intersection over the union defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

commonly known as the Jaccard index or the Jaccard similarity coefficient. To increase the speed of calculating the Jaccard similarity, minwise hahsing uses hash function to reduce the number of elements in set A and B, by taking advantage of the properties of minwise independent sets [?, pp. 3]. This property will be described below, as well as its application.

1.1.1 Min-wise Independency

Let $H : U \to [r]$ be a class of hashfunctions. For any set $X \subseteq [U]$ and any $x \in X$ and $h \in H$ chosen uniformly at random, it is considered minwise independent if

$$\Pr(h_{\min}(X) = h(x)) = \frac{1}{|X|} \tag{2}$$

where

$$h_{\min}(X) = \min\{\forall x \in X, h(x)\}$$

This means that all elements in X must have an equal probability of having the minimum value going through h. As seen in Eq. ??, this probability is reachable using universal hash functions.

1.1.2 Min-wise sketch

For two sets A and B it has been proven in [?] that Eq. 2 can be linked to the Jaccard similarity in Eq. 1 as

$$\Pr(h_{\min}(A) = h_{\min}(B)) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

For a random set S_1, a table of random hash functions $h_{\min,i}, i = 1, \ldots, n_h$ can be created such that

$$\hat{S_1} = \{h_{\min,1}(S_1), h_{\min,2}(S_1), \ldots, h_{\min,n_h}(S_1)\}$$

$\hat{S_1}$ is called to **min-wise sketch** of S_1. Given a set S_2 with the min-wise sketch $\hat{S_2}$, the similarity of S_1 and S_2 can then be defined by

$$J(A,B) = \frac{1}{n_h} \cdot \sum_{i=1}^{n_h} (h_{\min,i}(S_1) = h_{\min,i}(S_2)) \tag{4}$$

where

$$(h_{\min,i}(S_1) = h_{\min,i}(S_2)) = \begin{cases} 1, & h_{\min,i}(S_1) = h_{\min,i}(S_2) \\ 0, & otherwise \end{cases}$$

which is the **minwise similarity index**. The influence n_h will have on the error can be proven using Chernoff Bounds[1] [**?**] showing that the relation between n_h and ϵ, the error, is

$$\epsilon = O\left(\frac{1}{\sqrt{n_h}}\right) \tag{5}$$

Thus, an $n_h = 100$ should give approximately 10% error. The necessary number of hash functions for this error can however be halved by making a few slight modifications.

1.2 Max-wise hashing

The aforementioned modification of the min-wise hashing is one inspired by the method in the paper [**?**]. It is an extension of the minwise sketch where in addition to using the min-wise independent sets, the max-wise independent set is appended. Very literally, max-wise independent sets use the maximal hashvalue instead of the minimal hashvalues, such that a set X is said to be max-wise independent if

$$\Pr(h_{\max}(X) = h(x)) = \frac{1}{|X|}, h_{\max} = \max\{\forall x \in X, h(x)\} \tag{6}$$

for any $x \in X$. The Jaccard similarity measure for two sets A and B is

$$\Pr(h_{\max}(A) = h_{\max}(B)) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

and finally for a random set S_1, given a table of random $h_{\max,i}, i = 1, \ldots, n_h$, a sketch can be made, like so

$$\tilde{S_1} = \{h_{\max,1}(S_1), h_{\max,2}(S_1), \ldots, h_{\max,n_h}(S_1)\}$$

This sketch is called the **max-wise sketch**, and works almost like the minwise sketch. It is first when combining the max-wise- and min-wise sketches that one can find interesting similarity measures.

1.3 Max-wise and Min-wise similarity measures

Before considering the similarity measures, we shall define the sketches these will use. Given a set S_1, given two tables of random $h_{min,i}$ and $h_{max,i}$ of equal length where $i = 1, \ldots, n_h$, we have

$$\overline{S_1} = \{(h_{min,1}, h_{max,1}), \ldots, (h_{min,n_h}, h_{max,n_h})\} \tag{8}$$

which is called the **max-min-wise sketch**. There are two ways of using this sketch for similarity measures. One is the method described in [**?**], where the

[1] A probablistic method to find the exponentially decreasing bounds between two independent variates.

length of the sketch only needs to be half as long as the **min-wise sketch**, so that for $i = 1, \dots, n_h/2$

$$J(A,B) = \frac{1}{n_h}\left(\sum_{i=1}^{n_h/2} h_{\min,i}(A) = h_{\min,i}(B) + \sum_{i=1}^{n_h/2} h_{\max,i}(A) = h_{\max,i}(B)\right) \quad (9)$$

where

$$(h_{\max,i}(S_1) = h_{\max,i}(S_2)) = \begin{cases} 1, & h_{\max,i}(S_1) = h_{\max,i}(S_2) \\ 0, & otherwise \end{cases}$$

Let this method be called **Max-minwise halved similarity measure** (abbr. **MM**$\frac{1}{2}$). This method has been proven to be twice as fast as the min-wise hashing, without loss of precision [**?**]. It is also shown in Lemma 2 in [**?**] that for $i = 1, \dots, n_h/2$

$$\Pr((h_{\min,i}(A) = h_{\min,i}(B)) = 1 \mid (h_{\max,i}(A) = h_{\max,i}(B)) = 1) = \frac{|A \cap B| - 1}{|A \cup B| - 1}$$

Meaning that a collision between $h_{\min}$ and $h_{\max}$ is very unlikely.

Another method, which I developed in the course of this project uses the following similarity measure

$$J(A,B) = \frac{1}{n_h}\sum_{i=1}^{n_h}(h_{\min,i}(A) = h_{\min,i}(B) \vee h_{\max,i}(A) = h_{\max,i}(B)) \quad (10)$$

where

$$(h_{\min,i}(A) = h_{\min,i}(B)) \vee (h_{\max,i}(A) = h_{\max,i}(B)) = \begin{cases} 1, & h_{\min,i}(S_1) = h_{\min,i}(S_2) \\ 1, & h_{\max,i}(S_1) = h_{\max,i}(S_2) \\ 0, & otherwise \end{cases}$$

Let this method be called **Max-minwise similarity measure** (abbr. **MM**). The expected value of **MM** is Jaccard similarity coefficient by the following proof:

$$\begin{gathered} \frac{1}{n_h}\sum_{i=1}^{n_h}(h_{\min,i}(A) = h_{\min,i}(B) \vee h_{\max,i}(A) = h_{\max,i}(B)) = \\ \frac{1}{n_h}\sum_{i=1}^{n_h}(J(A,B) \vee J(A,B)) = J(A,B) \vee J(A,B) = J(A,B) \end{gathered} \quad (11)$$

the final three steps follow from Eq. 4 and Eq. 7. It follows that this method also calculates the Jaccard similarity.

As one may have noted, the difference between **Mm**$\frac{1}{2}$ and **Mm** is that the first runs only half as many times as the second for each comparison between two sets. The error of these functions can be deduced to

$$\epsilon = O\left(\frac{1}{\sqrt{2n_h}}\right) \quad (12)$$

by the error of the min-wise similarity measure in relation to n_h, meaning that $n_h = 50$ would give an error of approximately 10% .