Sketching Algorithms For Approximating Rank Correlations In Collaborative Filtering Systems

Yoram Bachrach[1], Ralf Herbrich[1], and Ely Porat[2]

[1] Microsoft Research Ltd., Cambridge, UK
[2] Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel

Abstract. Collaborative filtering (CF) shares information between users to provide each with recommendations. Previous work suggests using sketching techniques to handle massive data sets in CF systems, but only allows testing whether users have a high proportion of items they have both ranked. We show how to determine the *correlation* between the rankings of two users, using concise "sketches" of the rankings. The sketches allow approximating Kendall's Tau, a known rank correlation, with high accuracy ϵ and high confidence $1 - \delta$. The required sketch size is logarithmic in the confidence and polynomial in the accuracy.

1 Introduction

Recommender provide a user with recommendations regarding information items she is likely to find interesting. These systems compare user profiles to reference characteristics. Sometimes these characteristics are obtained from the content of the item (in the *content based* approach), and sometimes from information regarding the tastes of other users, in the *collaborative filtering (CF)* approach.

We consider a CF domain, where each user *ranks* the items she examined. Consider Alice, who asks the CF system to give a prediction for a certain item. The CF system must search for users who have ranked many of the items Alice has ranked. Then, the system should consider their rankings, and decide whether these users' tastes are similar to Alice's. A naive way to do this is to store the complete item lists and rankings for each user. However, this requires storing an enormous amount of data.The work [2] proposed a sketching technique for computing the *proportional intersection* (PI) of the ranked item lists. Rather than storing the full information they suggested *very concise* descriptions of ranked item lists, called *sketches*. Give a target accuracy $\epsilon > 0$ and a target confidence δ, their method returned an approximation $\hat{x}$ to the actual PI x, such that with probability of at least $1 - \delta$ the approximation is accurate enough, so $|\hat{x} - x| \leq \epsilon$. The major shortcoming of [2] is that *it did not allow computing a correlation grade between the rankings.* Even if there are many items ranked by both users, it is hard to construct a recommendation based solely on this information, as they may have given very different ratings these items. This work extends [2] and provides methods for computing the *correlation* between the rankings using *sketching techniques.* We construct an *extremely concise* representation of the user's item rankings, called a *rank correlation sketch*. Our sketches are designed to approximate Kendall's Tau [8], a well known rank correlation grade, while maintaining an only a small fraction of the information.

Consider Alice and Bob, who have each examined and ranked a set of n items, giving the item liked most has a rank of 1, the second best has a rank of 2, and so on until the worst item with the rank n. A known statistic to measure the correspondence between two rankings is Kendall's Tau [8]. Given the rankings of Alice and Bob, and given two items, A and B, we call the items a *concordant pair* if Alice and Bob agree on their order (i.e. if both Alice and Bob prefer A over B or if both prefer B to A). When Alice and Bob disagree on these items they are called a *discordant pair*. Given two rankings, we denote by n_c the number of concordant pairs, and by n_d the number of discordant pairs. Every pair is either concordant or discordant, so $n_c = n - n_d$.

Definition 1. *Kendall's Tau of r_a and r_b is:* $\tau_{r_a,r_b} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$

The total number of pairs is $\frac{1}{2}n(n-1)$, so $P(C) = \frac{n_c}{\frac{1}{2}n(n-1)}$ is the probability of a uniformly randomly chosen pair to be a concordant, and $P(D) = \frac{n_d}{\frac{1}{2}n(n-1)}$ is the probability for a discordant pair. Thus Kendall's Tau can be expressed as $\tau_{r_a,r_b} = p(C) - P(D) = P(C) - (1 - P(C)) = 2P(C) - 1$.

Consider the users of the CF system, $a_1, \ldots, a_m$. Each a_i has a ranking r_i of the items she had experience with. Under our model, we only maintain a sketch S_i of each ranking.

A sketching framework allows approximating Kendall's Tau τ_{r_i,r_j} between any two users, with a target confidence and accuracy.

Definition 2. *A rank correlation sketching framework with confidence δ and accuracy ϵ maintains only $S_1, S_2, \ldots, S_m$, and for any two users, a_i and a_j, allows computing τ_{r_i,r_j} with accuracy of at least ϵ and with confidence of at least $1 - \delta$. That is, the framework returns an approximation $\hat{\tau_{i,j}}$ for τ_{r_i,r_j} such that with probability of at least $1 - \delta$ we have $|\tau_{r_i,r_j} - \hat{\tau_{i,j}}| \leq \epsilon$.*

2 Sketches for Approximating Rank Correlation

Our sketching framework extends [2], so we first review that technique. Consider Alice and Bob, with the set C_1 of items that Alice has rated, and the set C_2 of items that Bob has rated, from the universe U of items, where $|C_1| = |C_2|$. Consider a sketch S_i that is the identity of a *single* item chosen uniformly at random from C_i. The probability of choosing the same item in S_1 and S_2 depends on $\frac{|C_1 \cap C_2|}{|C_1|}$, and is small. One insight comes from deciding to let the sketch S_i be the *minimal* item from C_i. If the minimal item in $C_1 \cup C_2$ is in $C_1 \cap C_2 = T$, we are guaranteed to find the item in $S_1 \cap S_2$. However, always using the *minimal* item always generates the same S_1, S_2. The methods in [2] overcome this by using min-wise independent hashes. Let H be a family of functions such that each $h \in H$ is a function $h : X \to Y$, where Y is completely ordered. We say

H is min-wise independent if, when randomly choosing $h \in H$, for any subset $C \subseteq X$, any $x \in C$ has an equal probability of being the minimal under h.

Definition 3. *H is min-wise independent, if for all $C \subseteq X$, for any $x \in C$, $Pr_{h \in H}[h(x) = min_{a \in C} h(a)] = \frac{1}{|C|}$.*

The work [7] constructs such families. The work in [2] uses them to build sketches for approximating the PI. In that work they use integers to define the identity of items in U (where $|U| = u$), so any subset of items $C \subseteq U$, is represented as a list of $|C|$ integers in $[u]$ ($[u]$ denoting $\{1, 2, \ldots, u\}$). They use a family H of min-wise independent functions from $[u]$ to $[n^2]$. Thus, although the domain is the huge universe of $[u]$ items, the hashed values are in the smaller range of $[n^2]$ items [3]. The methods of [2] consider users a_1, a_2, each with a list C_i of examined items, such that $|C_1| = |C_2|$. The sketches they propose approximate the PI between the two users, $p_{1,2} = \frac{|C_1 \cap C_2|}{|C_1|} = \frac{|C_1 \cap C_2|}{|C_2|}$. These sketches are based on randomly choosing hashes from H. Given $h \in H$, we can apply h on all the integers in C_1 and examine the minimal integer we get, $m_1^h = min_{x \in C_1} h(x)$. We can do the same to C_2 and examine $m_2^h = min_{x \in C_2} h(x)$. The following Lemma is proved in [2].

Lemma 1. $Pr_{h \in H}[m_1^h = m_2^h] = \frac{p_{1,2}}{2 - p_{1,2}}$.

We refer to the the sketches used by [2] as *item sketches*. They are defined as follows. Let $v_k = \langle h_1, h_2, \ldots, h_k \rangle$ be a tuple of k randomly chosen functions from the min-wise independent family H, and let C_i be the set of items that user a_i has examined. Denote the minimal item in C_i under h_j as $m_i^{h_j} = min_{x \in C_i} h_j(x)$.

Definition 4 (Item Sketches). *The H_k sketch of C_i, $S(C_i)$, is the list of minimal items in C_i under the k randomly chosen functions from h: $S^k(C_i) = (m_i^{h_1}, m_i^{h_2}, \ldots, m_i^{h_k})$.*

There are several key observations regarding item sketches. First, since H is min-wise independent, each sketch $S(C_i)$ on its own is a list of k random items from C_i (after applying a hash function on each item). Second, due to Lemma 1, randomly choosing a function $h \in H$ and testing whether $m_1^h = m_2^h$ is a Bernoulli trial, with success probability of $\alpha = \frac{p_{a,b}}{2 - p_{a,b}}$. We denote by X_i the random variable of the Bernoulli trial using hash h_i, so $X_i = 1$ if $m_a^{h_i} = m_b^{h_i}$, and $X_i = 0$ otherwise. Given an item sketch of k hashes, we get k such Bernoulli trials, $X_1, \ldots, X_k$, and can estimate $\alpha = \frac{p_{a,b}}{2 - p_{a,b}}$ as $\frac{\sum_{i=1}^{k} X_i}{k}$. Since $\alpha = \frac{p_{a,b}}{2 - p_{a,b}}$, we have $p_{a,b} = \frac{2\alpha}{1+\alpha}$, so given an estimate $\hat{\alpha}$ for α, we can estimate $p_{a,b}$ as $\hat{p_{a,b}} = \frac{2\hat{\alpha}}{1+\hat{\alpha}}$. The work [2] shows that to approximate the PI $p_{a,b}$ within accuracy ϵ and confidence $1 - \delta$, it is enough to use $k = \frac{\ln \frac{2}{\delta}}{2 \frac{\epsilon^2}{9}}$ hashes. The methods in [2] do not show how to compute the correlation between the rankings.

[3] The methods of [2] build on the results of [7], which show that using a range of n^2 integers mitigates the effect of collisions in the hashed values. Thus, the probability of two different items in $[u]$ to be mapped to the same value after applying the hash (a collision) is very small.

2.1 Rank Correlation Sketches

We now describe our method for constructing a rank correlation sketching framework. We first return to Alice and Bob. Now suppose we have a set of items I that *both* Alice and Bob have ranked, from the universe U of items. Denote by n the size of I, so $|I| = n$. We are interested in approximating τ_{r_a,r_b}, Kendall's Tau rank correlation between Alice's ranking of the items in I and Bob's ranking. We denote by n_c the number of concordant pairs, and by n_d the number of discordant pairs, so $\tau_{r_a,r_b} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$. As noted in Section 1, the probability $P(C) = \frac{n_c}{\frac{1}{2}n(n-1)}$ is closely relate to Kendall's Tau, and $\tau_{r_a,r_b} = 2P(C) - 1$.

Lemma 2 (Approximating $P(C)$ and Kendall's Tau). *Approximating $P(C)$ with accuracy $\frac{\epsilon}{2}$ gives an approximation to Kendall's Tau with accuracy ϵ.*

Proof. We use the approximation $P(\hat{C})$ for $P(C)$ to approximate Kendall's Tau. Our approximation for τ_{r_a,r_b} is $\tau_{\hat{r_a,r_b}} = 2P(\hat{C}) - 1$. If our error in our estimation of $P(C)$ is at most $\frac{\epsilon}{2}$, we have $|P(C) - P(\hat{C})| \leq \frac{\epsilon}{2}$, so $|\tau_{r_a,r_b} - \tau_{\hat{r_a,r_b}}| = |2P(C) - 1 - (2P(\hat{C}) - 1)| = |2(P(C) - P(\hat{C}))| \leq 2 \cdot \frac{\epsilon}{2} = \epsilon$. Thus, to approximate Kendall's Tau with accuracy ϵ it is enough to approximate $P(C)$ with accuracy $\frac{\epsilon}{2}$.

We now consider a pair of items chosen uniformly at random from I, $x, y \in I$. Given Alice's and Bob's rankings, we can test whether this is a concordant pair. This is a Bernoulli trial, with a success probability of $P(C)$. We define the random variable of this Bernoulli trial as: $X_1 = \begin{cases} 1 & \text{if } x, y \text{ is a concordant pair} \\ 0 & \text{if } x, y \text{ is a discordant pair} \end{cases}$

Given k_p such pairs, we have a sequence of k_p such Bernoulli trials, $X_1, \ldots, X_{k_p}$. Let X be the number of successes in this series of Bernoulli trials, $X = \sum_{j=1}^{k_p} X_j$. We have chosen the pairs uniformly at random, so the X_is are identical but independent. Thus X has the Binomial distribution $X \sim B(k, \alpha)$, and the *maximum likelihood estimator* for $P(C)$ is $P(\hat{C}) = \frac{X}{k_p}$. We now derive the required number of random item pairs required to approximate $P(C)$ with accuracy ϵ_c and confidence δ_c. To achieve the desired accuracy and confidence, the number of sampled pairs, k_p, must be large enough. We find the appropriate k_p by using Hoeffding's inequality [6].

Theorem 1 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent random variables, where all X_i are bounded so that $X_i \in [a_i, b_i]$, and let $X = \sum_{i=1}^{n} X_i$. Then the following inequality holds:* $\Pr(|X - \mathrm{E}[X]| \geq n\epsilon) \leq 2\exp\left(-\frac{2\,n^2\,\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$.

Let $X_1, \ldots, X_{k_p}$ be the series k_p of Bernoulli trials, as defined above. Again, let $X = \sum_{j=1}^{k_p} X_j$, and take $P(\hat{C}) = \frac{X}{k_p}$ as an estimator for $P(C)$. All X_i are either 0 or 1 (so they are bounded between these values), and $\mathrm{E}[X] = k_p \cdot P(C)$. Thus, the following holds: $\Pr(|X - k_p P(C)| \geq k\epsilon_c) \leq 2e^{-2\,k_p\,\epsilon_c^2}$. Therefore the following also holds: $\Pr(|P(\hat{C}) - P(C)| \geq \epsilon_c) \leq 2e^{-2\,k_p\,\epsilon_c^2}$. We now extract the number of pairs required so that this probability is below some required confidence level δ_c.

Theorem 2 (Pair Samples for Approximating $P(C)$). *A confidence interval for $P(C)$ is $[P(\hat{C}) - \epsilon_c, P(\hat{C}) + \epsilon_c]$. This interval holds the correct $P(C)$ with probability of at least $1 - \delta_c$. The required number of pair samples to perform this is $k_c \ = \frac{\ln \frac{2}{\delta_c}}{2\,\epsilon_c^2}$.*

Proof. We use Hoeffding's inequality to bound the error below our target confidence level δ_c, and get: $Pr(|P(\hat{C}) - P(C)| \geq \epsilon_c) \leq 2\,e^{-2\,k_c\,\epsilon_c^2} \leq \delta_c$. We extract ϵ_c and k_c: $-2\,k_c\,\epsilon_c^2 \leq \ln \frac{\delta_c}{2}$. Equivalently: $\epsilon_c^2 \geq -\frac{\ln \frac{\delta_c}{2}}{2\,k}$. Finally we get the following: $\epsilon_c \geq \sqrt{\frac{1}{2k_c} \ln \frac{2}{\delta_c}}$ and $k_c \geq \frac{\ln \frac{2}{\delta_C}}{2\,\epsilon_c^2}$.

The required number of pairs in Theorem 2 considered approximating $P(C)$ and not Kendall's Tau. However, due to Lemma 2 we get the following corollary.

Corollary 1 (Pair Samples for Approximating Kendall's Tau). *The following is an approximation for Kendall's Tau: $\tau_{\hat{r_a,r_b}} = 2P(\hat{C}) - 1$. In order for it to have accuracy ϵ_t and confidence δ_t the required number of random pairs samples is $k_t \ = \frac{2\ln \frac{2}{\delta_t}}{\epsilon_t^2}$.*

2.2 From Item Sketches To Rank Correlation Sketches

We now augment the sketches of [2] to approximate Kendall's Tau. The PI sketches of [2] approximate the PI. When the CF system attempts to provide Alice (with items C_a) with a recommendation, it filters out users who do not have a high enough PI with her, so only users with a PI exceeding a threshold, p^*, remain. Consider a candidate, Bob (with item set C_b), where the PI of Alice and Bob is $p_{a,b}$. By definition of the PI, $p_{a,b} = \frac{|C_a \cap C_b|}{|C_a|} = \frac{|C_a \cap C_b|}{|C_b|}$ [4], and since Bob has passed the filtering stage we have $p_{a,b} \geq p^*$. The item sketch from Definition 4 randomly chooses k hashes from H, and lists the minimal items under these k hashes. By definition of H as a min-wise independent family, for any user's set of items C, any item has an equal probability of being minimal under the hash, so $Pr_{h \in H}[h(x) = min_{a \in C} h(a)] = \frac{1}{|C|}$. Let $h \in H$ be a randomly chosen hash function from H. We denote the minimal item in C_i under h as $m_i^h = min_{x \in C_i} h(x)$. We show that if Alice and Bob have a PI of at least p^*, the probability of having the same value at each sketch location is at least $\frac{p^*}{2-p^*}$.

Lemma 3 (Probability Of The Same Item Appearing In Two Sketches). *Let Alice and Bob be two users with a PI of at least p^*, Alice with item set C_a and Bob with item set C_b. Then $Pr_{h \in H}[m_a^h = m_b^h] \geq \frac{p^*}{2-p^*}$ (i.e. the probability of Alice and Bob having* same *minimal item under h is at least $\frac{p^*}{2-p^*}$).*

Proof. We denote the PI of Alice and Bob as $p_{a,b} \geq p^*$. Due to Lemma 1 we have $Pr_{h \in H}[m_a^h = m_b^h] = \frac{p_{a,b}}{2-p_{a,b}}$, and since $f(x) = \frac{x}{2-x}$ is monotonically increasing in the domain $[0, 1]$ we have $Pr_{h \in H}[m_a^h = m_b^h] \geq \frac{p^*}{2-p^*}$.

[4] Note that we are still assuming the same size of item set per user.

Consider Alice and Bob, with a PI of at least p^*. Lemma 3 states that any location i has a probability of at least $p_s = Pr_{h_i \in H}[m_a^{h_i} = m_b^{h_i}] \geq \frac{p^*}{2-p^*}$ to contain the same value in Alice's sketch and in Bob's sketch. Since the range of the hashes in H is $[n^2]$ (where n is the number of items examined by each user), having the same minimal item under h, $m_a^{h_i} = m_b^{h_i}$, indicates with high probability that this is the *same* item, so $|\{x \in C_a | h_i(x) = m_a^{h_i}\}| = |\{y \in C_b | h_i(y) = m_b^{h_i}\}| = 1$, and both C_a and C_b contain only one item x (so $x \in C_a$ and $x \in C_b$) such that $h_i(x) = m_a^{h_i} = m_b^{h_i}$.

Let h_i be the hash for the i'th location in the item sketch. The augmenting part of the sketch includes the rank of the item that is minimal under h. We denote the ranking of user a over the items in C_a as r_a, so r_a maps items from C_a to their rank in $[n]$ (where $|C_a| = n$). Thus $r_a : C_a \rightarrow [n]$ is reversible. We randomly choose a hash for each sketch location. Given the hash h_i for location i, we consider the items who are minimal under h_i, i.e. $M = \{x \in C_a | h_i(x) = m_a^{h_i}\}$. If $|M| = 1$ we denote $M = \{m\}$, and denote $g_a^i = r_a(m)$. If $|M| \geq 1$, which occurs with a very low probability, we denote m' to be the minimal item in M (under the original ordering, not under h_i), and denote $g_a^i = r_a(m')$. The sketch for user a in the i'th location contains the minimal item in C_a under h_i, and its ranking in user a's eyes.

Definition 5 (Rank Correlation Sketches). *The H_k rank correlation sketch of C_a, $S^k(C_a)$, contains the both the item sketch and the rank sketch. As before, the item sketch is just the list of minimal items in C_a under the k randomly chosen hashes, so $S^k_{items}(C_a) = (m_a^{h_1}, m_a^{h_2}, \dots, m_a^{h_k})$, and the rank sketch contains the ranks of these items, so $S^k_{ranks}(C_a) = (g_a^1, \dots g_a^n)$. The rank correlation sketch is simply the concatenation of these two sketches.*

Consider two locations i and j where the item sketch for both Alice and Bob is the same, i.e where $m_a^{h_i} = m_b^{h_i}$ and $m_a^{h_j} = m_b^{h_j}$. Each such location is called a *sketch collision*. Given two such collisions, with high probability the ranking sketch at these two locations refers to the same items, i.e. there are two items x, y such that $x, y \in C_a$ and $x, y \in C_b$, and such that $g_a^i = r_a(x), g_b^i = r_b(x), g_a^j = r_a(y), g_b^j = r_b(y)$. Since any item has an equal probability to be minimal under a random hash $h \in H$ (as H is min-wise independent), the sketches in these locations provide us with Alice's and Bob's rankings for a pair of items chosen uniformly at random from $C_a \cap C_b$. Corollary 1 gives the required number of pairs to approximate Kendall's Tau, but each pair requires two independent collisions [5]. Thus, approximating Kendall's Tau is reduced to finding a sketch that would have the required number of collisions with high probability.

2.3 Collisions And Sketch Size

Consider Alice, who seeks a recommendation from the CF system. The CF system has a PI threshold p^*, and selects only candidates who have a higher PI

[5] Notice that given m sketch collisions we can generate $\frac{m(m-1)}{2}$ pairs, but these pairs would not be independent.

with her. As shown in [2], in order to compute the PI with accuracy ϵ_i and confidence δ_p, it is enough to use a sketch based on $k_p \ = \frac{\ln\frac{2}{\delta_p}}{2\,\frac{\epsilon_p^2}{9}}$ hashes. After filtering out candidates with too low a PI, the CF system remains with candidates, and computes Kendall's Tau for each of them, with accuracy ϵ_t and confidence $1-\delta_t$.

Lemma 3 shows that the probability of a collision in each location is at least $p = \frac{p^*}{2-p^*}$. Thus each location is a Bernoulli trial, with success probability of at least p (success being a collision). Theorem 1 shows that approximating Kendall's Tau with accuracy ϵ_t and confidence δ_t requires $2k_t \ = \frac{4\ln\frac{2}{\delta_t}}{\epsilon_t^2}$ sketch collisions. We determine the size of the sketch needed to have such a required number of collisions with probability of at least $1-\delta_c$. Given a sketch based on m hashes, the number of collisions X has the Binomial distribution with parameters m, p. We require $k = 2k_t$ collisions, and thus are interested in the cumulative distribution function $F(k,m,p) = P(X \leq k) = \sum i = 0^k \binom{m}{i} p^i (1-p)^{n-i}$. We find a sketch size m that is high enough that $F(k,m,p)$ is below our confidence level δ_c, using the following result from [3]:

Theorem 3 (Binomial Distribution Tail Bound). $F(k,m,p) \leq \exp\left(-2\frac{(mp-k)^2}{n}\right)$

Theorem 4 (Rank Correlation Sketch Size). *A rank correlation sketching framework for users with PI of at least* p^*, *where* $p = \frac{p^*}{2-p^*}$, *requires sketch size of* $m \geq \frac{k}{p} + \frac{\ln\frac{1}{\delta_c}}{4p^2}(1+3\sqrt{k})$, *where* $k = 2k_t \ = \frac{4\ln\frac{2}{\delta_t}}{\epsilon_t^2}$.

Proof. We require a sketch size m such that $F(k,m,p) \leq \exp\left(-2\frac{(mp-k)^2}{n}\right) \leq \delta_c$. Thus we require $\frac{-2(mp-k)^2}{n} \leq \ln\delta_c$, or that $\frac{(mp-k)^2}{m} \geq \frac{\ln\frac{1}{\delta_c}}{2}$. We denote $d = \frac{\ln\frac{1}{\delta_c}}{2}$. The requirement is thus that $p^2m^2 + (-2pk-d)m + k^2 \geq 0$. Solving the quadratic equation (and taking the bigger solution) we get $m \geq \frac{2pk+d+\sqrt{4pkd+d^2}}{2p^2}$ or that $m \geq \frac{k}{p} + \frac{d}{2p^2} + \frac{\sqrt{4pkd+d^2}}{2p^2}$. An even strong requirement is that $m \geq \frac{k}{p} + \frac{d}{2p^2}(1+\sqrt{4k+1})$, or even that $m \geq \frac{k}{p} + \frac{d}{2p^2}(1+\sqrt{9k}) = \frac{k}{p} + \frac{d}{2p^2}(1+3\sqrt{k})$. We finally get that the requirement is $m \geq \frac{k}{p} + \frac{\ln\frac{1}{\delta_c}}{4p^2}(1+3\sqrt{k})$.

Thus the sketch size is polynomial in the accuracy, and logarithmic in the confidence [6].

3 Related Work

There are many examples of CF systems, such as GroupLens [9] and Ringo [11]. [9] uses the Pearson correlation, while [11] uses other measures. This paper tackles the problem of handling the massive data sets in CF systems. We suggested

[6] There are two confidence levels, δ_t the maximal probability of mis-approximating Kendall's Tau, and δ_c, the maximal probability of not having enough sketch collisions. From the union bound, the probability of having a bad approximation is at most $\delta_c + \delta_t$, and the sketch size is logarithmic in both.

sketching to approximate rank correlations. One example of a sketching technique is [4]. We extend [2] to compute rank correlations, using a min-wise independent family of hashes. Such families were treated in [7]. Our methods fits in the Locally Sensitive Hashing (LSH) [5] framework, but is specialized for CF systems. Similar approach are Random Projections [1] and Semantic/Spectral Hashing [10, 12].

4 Conclusion

A challenge in CF systems is handling huge amounts of information. We have suggested a sketching approach to *approximate* the rank correlation with a given accuracy and confidence. The sketch size is logarithmic in the confidence, and polynomial in the accuracy. There are many directions for future research. Our methods only allow computing Kendall's Tau and not other rank correlations, such as Spearman's Rho. Also, we assume a complete ranking over items, and do not allow for ties. Another shortcoming of our analysis here is that it is only theoretical. It would be desirable to test these methods on real data sets.

References

1. Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *JCSS*, 66, 2003.
2. Yoram Bachrach, Ely Porat, and Jeffrey S. Rosenschein. Sketching techniques for collaborative filtering. In *IJCAI 2009*, Pasadena, California, July 2009. To appear.
3. Kai Lai Chung. *Elementary Probability Theory with Stochastic Processes*. Springer-Verlag, 1974.
4. Joan Feigenbaum, Sampath Kannan, Martin Strauss, and Mahesh Viswanathan. An approximate L1-difference algorithm for massive data streams. *SIAM J. Comput*, 32(1):131–151, 2002.
5. Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *VLDB: International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers, 1999.
6. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
7. Piotr Indyk. A small approximately min-wise independent family of hash functions. *Journal of Algorithms*, 38(1):84–90, January 2001.
8. Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.
9. P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 175–186. ACM, 1994.
10. Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, December 2008.
11. Upendra Shardan and Pattie Maes. Social information filtering: Algorithms for automating "word of mouth". In *ACM CHI'95*, volume 1, pages 210–217, 1995.
12. Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *Advances in Neural Processing Systems*, 2008.