Fast Asynchronous Anti-TrustRank for Web Spam Detection

Joyce Jiyoung Whang\textsuperscript{1}, Yeonsung Jung\textsuperscript{1}, Inderjit S. Dhillon\textsuperscript{2}, Seonggoo Kang\textsuperscript{3}, and Jungmin Lee\textsuperscript{3}

\textsuperscript{1}Sungkyunkwan University (SKKU), \textsuperscript{2}The University of Texas at Austin, \textsuperscript{3}Naver Corporation.

ACM International Conference on Web Search and Data Mining (WSDM) Workshop on MIS2: Misinformation and Misbehavior Mining on the Web, 2018

Main Contributions

\begin{itemize}
    \item Asynchronous Anti-TrustRank algorithms
    \item Significantly reduce the number of arithmetic operations compared to the traditional synchronous Anti-TrustRank algorithm
    \item Without degrading the performance in detecting Web spams
    \item Convergence of the asynchronous Anti-TrustRank algorithms
    \item Experiments on a real-world Web graph indexed by NAVER which is the most popular search engine in Korea.
\end{itemize}

Notation

\begin{itemize}
    \item \(G = (V, E')\): a graph with reverse edges, i.e., if an edge \(\{i, j\} \in E\) then \(\{j, i\} \in E'\). Also, let \(A\) denote the adjacency matrix of \(G\).
    \item \(P \equiv D^{-1}A\) (\(D\) is the degree diagonal matrix)
    \item \(Q\): the set of incoming neighbors of node \(i\) on \(G\)
    \item \(T_i\): the set of outgoing neighbors of node \(i\) on \(G\)
    \item \(x\): a vector of the ATR scores, \(\alpha\): a vector of the residuals
    \item \(e_i\): a vector with ones for the positions of the seed spam documents and zeros for other positions
\end{itemize}

Anti-TrustRank

\begin{itemize}
    \item Spam pages are likely to be referred by other spam pages.
    \item Documents with high Anti-TrustRank (ATR) score \(\rightarrow\) spam pages
    \item From spam seeds, the ATR scores are propagated to incoming neighbors of the nodes so that the documents having links to the spam documents end up with having high ATR scores.
\end{itemize}

* V. Krishnan et al., Web spam detection with anti-trust rank. AIRWeb, 2006.

Algorithms

\begin{itemize}
    \item Synchronous Anti-TrustRank (SYNC ATR)
        \begin{itemize}
            \item The scores are updated after all the nodes re-compute the scores.
        \end{itemize}
    \item Asynchronous Anti-TrustRank (ASYNC ATR)
        \begin{itemize}
            \item \texttt{worklist}: a set of nodes whose ATR scores need to be updated.
        \end{itemize}
    \item Residual-based Asynchronous Anti-TrustRank (RASYNC ATR)
        \begin{itemize}
            \item new ATR = current ATR + current residual (explicitly maintain the residual of each node)
            \item Filtering out unnecessary work in the \texttt{worklist}.
        \end{itemize}
\end{itemize}

Experimental Results

\begin{itemize}
    \item Real-world Web graph from NAVER corporation
        \begin{itemize}
            \item 584,092 documents and 2,470,557 edges
            \item 437,386 (74.88\%) normal docs and 45,641 (7.81\%) spam docs
            \item 101,065 (17.30\%) documents are unlabeled.
        \end{itemize}
    \item From spam seeds, the ATR scores are propagated to incoming neighbors of the nodes so that the documents having links to the spam documents end up with having high ATR scores.
\end{itemize}

South Korea’s No. 1 Search Portal

\begin{itemize}
    \item The most popular search port in South Korea
    \item Search query share in Korea: 74.7% (As of June 2016)
    \item Daily visitors of mobile HTML2: 27M. Registered Users: 42M
\end{itemize}

Most of the retrieved documents are correctly classified into spam.

\begin{itemize}
    \item \(|\mathcal{L}| = p|V|\) where \(\mathcal{L}\) denotes the set of labeled documents
    \item Pick top \(q|S|\) documents where \(S\) denotes the set of spam seeds
\end{itemize}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
\(q = 1\) & \(q = 3\) & \(q = 5\) \\
\hline
\hline
\(p = 0.01\) & \begin{tabular}{c}
spam docs 1,367 (100\%) \\
normal docs 0 (0\%)
\end{tabular} & \begin{tabular}{c}
spam docs 4,099 (99.951\%) \\
normal docs 0 (0\%)
\end{tabular} \\
unlabeled docs 0 (0\%) & 0 (0\%) & 0 (0\%) & 0 (0\%)
\hline
\(p = 0.02\) & \begin{tabular}{c}
spam docs 3,083 (100\%) \\
normal docs 0 (0\%)
\end{tabular} & \begin{tabular}{c}
spam docs 9,113 (98.530\%) \\
normal docs 0 (0\%)
\end{tabular} \\
unlabeled docs 0 (0\%) & 2 (0.049\%) & 2 (0.029\%)
\hline
\(p = 0.03\) & \begin{tabular}{c}
spam docs 3,910 (100\%) \\
normal docs 0 (0\%)
\end{tabular} & \begin{tabular}{c}
spam docs 11,593 (98.832\%) \\
normal docs 0 (0\%)
\end{tabular} \\
unlabeled docs 0 (0\%) & 10 (0.912\%) & 10 (0.547\%)
\hline
\end{tabular}
\caption{Accuracy of the retrieved documents}
\end{table}

The asynchronous algorithms, ASYNC and RASYNC, make much fewer ATR updates than the synchronous algorithm, SYNC.

RASYNC significantly reduces the number of arithmetic computations.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\(p\) & \(\epsilon\) & SYNC & ASYNC & RASYNC \\
\hline
\hline
0.01 & 10\(^{-8}\) & \begin{tabular}{c}
No. of ATR updates 2,336,368 \\
No. of arithmetic 24,442,660
\end{tabular} & \begin{tabular}{c}
20,361 \\
8,244,970
\end{tabular} \\
& 10\(^{-12}\) & 2,920,460 & 39,483 & 39,483 \\
& 0.03 & \begin{tabular}{c}
No. of ATR updates 30,566,040 \\
No. of arithmetic 30,553,325
\end{tabular} & 21,871,604 & 3,284,207
\hline
\hline
0.01 & 10\(^{-8}\) & \begin{tabular}{c}
No. of ATR updates 2,336,368 \\
No. of arithmetic 24,452,832
\end{tabular} & \begin{tabular}{c}
20,628 \\
10,716,065
\end{tabular} \\
& 10\(^{-12}\) & 2,920,460 & 39,804 & 39,804 \\
& 0.03 & \begin{tabular}{c}
No. of ATR updates 30,566,040 \\
No. of arithmetic 30,566,040
\end{tabular} & 25,817,600 & 3,932,405
\hline
\end{tabular}
\caption{No. of ATR updates and arithmetic operations}
\end{table}