

Diplomarbeit “ A Statistical Approach towards Spam Filtering ”

Dieses Dokument gibt den Rahmen der Diplomarbeit von Christian Wassmer im WS 2004/2005 vor. Abweichungen oder Änderungen sind in gegenseitiger Absprache möglich. Die vorgegebenen Zeiten sind als ungefähre Richtlinien gedacht.

Arbeitsplan:

Gegenstand der Arbeit ist die statistische Analyse von (Spam-)Nachrichten zur Verbesserung der Spam-Erkennungsrate. Die gewonnenen Erkenntnisse sollen einerseits dazu dienen, neue Spam-Filter für das Spamoto-System zu entwickeln, und andererseits nützliche Informationen über Spam sowohl client- als auch serverseitig zur Verfügung stellen. Desweiteren sollen die Ergebnisse die Entwicklung einer neuen Trust-Komponente unterstützen.

- 1) Ausarbeitung eines Konzeptes, um möglichst viele Spam-Nachrichten zu erhalten (Spamato-Archiv, spamarchive.org, Honeypots, ...) und diese nach verschiedenen Kriterien möglichst automatisch zu verarbeiten. Anhand welcher Kriterien (Google, DNS, RTT, Random Noise, ...) sollen Nachrichten untersucht werden? Wie können die Ergebnisse vernünftig präsentiert werden? [3W]
- 2) Einarbeitung in Spamato. [1W]
- 3) Entwicklung einer oder mehrere Filter für Spamato unter Berücksichtigung der ermittelten Ergebnisse. Unter anderem soll das Problem der „Multi-Domain“-Nachrichten (Nachrichten, in denen mehrere - „gute“ und „böse“ - Domains vorhanden sind) behandelt werden. [2W]
- 4) Überarbeitung der client- und serverseitigen Spamato-Statistiken. [2W]
- 5) Entwicklung einer neuen Trust-Komponente für Spamato. Können aus den an das Spamato-System gemeldeten Nachrichten (eindeutige) Benutzergruppen erstellt werden (wie sehen diese aus)? Können diese verwendet werden, um die vorhandene Trust-Komponente zu verbessern bzw. eine neue zu erstellen? [4W]
- 6) Zum Abschluss werden zwei Dokumente erstellt: [5W]
 - Ein Bericht (30 bis 50 Seiten, Sprache wählbar), welcher über die Arbeit und die Resultate Auskunft gibt. Dieser Bericht soll unter anderem auch eine kritische Beurteilung der eigenen Arbeit enthalten.
 - Einen Forschungsbericht (10 Seiten, in Englisch), welcher die Arbeit und die Resultate kompakter und unter einem wissenschaftlichen Gesichtspunkt darstellt.

Allgemeines:

- Selbstständiges Arbeiten ist Voraussetzung. Im TIK steht ein Arbeitsplatz zur Verfügung. Es besteht jedoch auch die Möglichkeit, zu Hause zu arbeiten.
- Während und nach der Arbeit ist eine Zwischen- (nach 2 Monaten) bzw. Abschluss-Präsentation vorgesehen.
- Monatlich einen „Monthly Report“ mit einer Übersicht der Arbeiten des abgelaufenen Monats.
- Pro Woche findet mindestens ein Treffen mit Keno statt.

Kontaktpersonen:

1. Keno Albrecht kenoa@tik.ee.ethz.ch
2. Roger Wattenhofer wattenhofer@tik.ee.ethz.ch