



iSquare Spider

Die folgenden Features gelten für alle iSquare Spider Versionen:

Feature	Beschreibung
webbasiertes Management (Administratoren)	<ul style="list-style-type: none">• Sichten aller gefundenen Beiträge eines Forums• Statusüberprüfung• Informationen zu:<ul style="list-style-type: none">• freier Speicher• verarbeitete URLs pro Minute• Größe und Anzahl der betrachteten HTML-Seiten• Anzahl der gefundenen Artikel• Aktivitätszeitplan• Dauer eines kompletten Suchvorgangs• Fehlertoleranz (mehrmalige Downloadversuche bei Netzwerkfehlern)• Anzahl der erfolgreichen/verwehrt Downloads pro Forum• Überwachung der Parallelität• Historie der vergangenen Zustände alter Spiderläufe• Mögliche Deaktivierung einzelner Quellen für einen Spiderlauf• Anpassbarkeit der Strukturbeschreibung für einzelne Quellen• Sichtung von Fehlermeldungen
Monitoring	Zur Überwachung der Anwendung ist die Genauigkeit des Fehlerreportings konfigurierbar.
Automatischer Restart	Im Fehlerfall wird die Anwendung automatisch neu gestartet und der Administrator wird per Email benachrichtigt
Parallelität (konfigurierbar)	Jeder Spider arbeitet beliebig viele Quellen parallel ab.
Löschen alter Datensätze	Datensätze können nach einem festzulegenden Zeitraum automatisch gelöscht werden



Feature	Beschreibung
Zentrale Haltung der Profile	Alle Profile können zentral in einer Datenbank gehalten werden. Jeder Spider liest die ihn betreffenden Konfigurationen von dort aus.
Überlastschutz pro Site	Um die Gefahr des Aussperrens durch den Betreiber des Hosts zu verringern, lassen sich für die Zugriffe auf ein und denselben Host die Zeit-Intervalle einstellen (min. bis max. Millisekunden). Dabei wird definiert, wann der nächste Zugriff auf denselben Host wieder erfolgen darf. Pro Anfrage wird dann ein Zufallswert aus dem Intervall ausgewählt.
Spracherkennung	Es wird erkannt, in welcher Sprache ein Artikel verfasst ist. (Deutsch/Englisch)
Extraktion von Artikel-Attributen	Neben dem eigentlichen Artikel können, frei definierbar, weitere Attribute aus dem Artikel extrahiert werden, wie z.B. Autor, Titel etc. Dies kann manuell erfolgen oder automatisch anhand von Trainingsdaten. Dadurch können z.B. zu alte Artikel vor der weiteren Verarbeitung verworfen werden.
Automatisierung der Erstellung von Profilen	Für die Extraktion von Artikel-Attributen können die Profile automatisch erzeugt werden. Hierdurch ergeben sich erhebliche Zeitersparnisse bei der Profilerstellung.



Spezielle Features

1. **iSquare Spider HTTP**
2. **iSquare Spider Forum**
3. **iSquare Spider RSS**
4. **iSquare Spider Suchmaschine**
5. **iSquare Spider Newsgruppe**
6. **iSquare Spider Email**

1. iSquare Spider HTTP

Feature	Beschreibung
Bearbeitungsliste	Verarbeitung einer definierten Liste von Websites.
Spidertiefe	Die Spidertiefe (Anzahl der Ebenen einer Website) ist frei konfigurierbar.
Spidern von Links einer Website	Links einer Webseite können explizit ein- oder ausgeschlossen werden
Datumserkennung	Das Datum wird erkannt und in ein einheitliches Format gebracht. Pro Site kann auch manuell ein sehr spezielles Datumsformat definiert werden.
Profil-Definition mit HTML-Mustern	Für die Erkennung von Artikeln als auch von Artikel-Attributen können HTML-Muster eingesetzt werden. Sollten diese Muster nicht ausreichen, können reguläre Ausdrücke verwendet werden.



2. iSquare Spider Forum

Der iSquare Spider Forum identifiziert aktuelle Einzelbeiträge in Internetforen, extrahiert Metadaten der Beiträge und speichert diese ab.

Feature	Beschreibung
Strukturidentifizierung des Forums	Identifizierung von Threads und Beiträgen Die Beiträge werden als eigenständige Dokumente erkannt und behandelt.
Autorenerkennung	Der Name/Nickname des Autors wird erkannt. Wenn möglich wird das Geschlecht anhand einer Wissensbasis vom Namen abgeleitet.
Titelerkennung	Die Betreffzeile wird erkannt.
Textextraktion	Der Text des Beitrages wird ohne störende HTML-Elemente erfasst.
Intelligente Aktualisierung	Der iSquare Spider Forum untersucht nur die seit dem letzten Durchlauf hinzugekommenen Dokumente. Die Treffer können daher schneller gefunden werden, und es werden keine Dokumente doppelt erfasst.
Profil-Definition mit HTML-Mustern	Für die Erkennung von Artikeln als auch Artikel-Attributen können HTML-Muster eingesetzt werden. Sollten diese Muster nicht ausreichen, können reguläre Ausdrücke verwendet werden.



3. iSquareSpider RSS

RSS ist ein Dateiformat für den XML-basierten Austausch von Nachrichten aller Art. Es geht bei RSS-Formaten immer darum, Informationen strukturiert abzulegen und sie für die automatisierte Verarbeitung durch RSS-Leseprogramme bereitzustellen. RSS wurde geschaffen, um Nachrichten von Internetportalen zu verbreiten, und hat sich inzwischen zu einem weit verbreiteten Standard für den automatisierten Austausch von Nachrichten und menschlicher Kommunikation z.B. in Weblogs und Diskussionsforen entwickelt.

Feature	Beschreibung
Quellen	Dieser Spider durchsucht die URLs, die in den RSS-Feeds angegeben sind, derzeit ca.10.000 Feeds in 20 Kategorien. Jedem Feed sind wiederum x Artikel zugeordnet.
Titelerkennung	Die Betreffzeile wird erkannt.
Datumserkennung	Das Datum wird erkannt und in ein einheitliches Format gebracht.
URL-Erkennung	Extraktion von im Text enthaltenen URLs.
Master-Slave-Konfiguration	Bei sehr vielen zu beobachtenden Feeds, kann ein Master die Feeds an mehrere Slaves verteilen. Die Bearbeitungsgeschwindigkeit ist damit frei skalierbar.
Zentrale Log-Erfassung	Die Slaves melden ihre Log-Nachrichten zentral an den Master.



4. iSquare Spider Suchmaschine

Der Spider Suchmaschine übergibt ein oder mehrere Begriffe an mehrere Suchmaschinen und speichert die dort gefundenen Treffer ab.

Feature	Beschreibung
Anzahl der Suchmaschinen	Es können mehrere Suchmaschinen gleichzeitig abgefragt werden. Zur Zeit Google, Yahoo, Altavista, u.a.
Anzahl der Suchbegriffe	Die Anzahl der Suchbegriffe ist unbegrenzt.
Mandanten-fähigkeit	Es ist möglich, gleichzeitig mehrere Suchbegriffe bei mehreren Suchmaschinen zu verschiedenen Themen abzufragen. Die Suchmaschinentreffer werden dennoch themenspezifisch gespeichert.
Intelligente Aktualisierung	Der Spider untersucht nur die seit dem letzten Durchlauf hinzugekommenen Artikel.
Datumserkennung	Das Datum wird erkannt und in ein einheitliches Format gebracht.
Profil-Definition mit HTML-Mustern	Für die Erkennung von Artikeln als auch von Artikel-Attributen können HTML-Muster eingesetzt werden. Sollten diese Muster nicht ausreichen, können reguläre Ausdrücke verwendet werden.



5. iSquare Spider Newsgruppe

Dieser Spider beobachtet ca. 2000 deutschsprachige Newsgruppen.

Feature	Beschreibung
Autorenerkennung	Der Name/Nickname des Autors wird erkannt. Wenn möglich wird das Geschlecht anhand einer Wissensbasis ebenfalls vom Namen abgeleitet.
Titelerkennung	Die Betreffzeile wird erkannt.
Datumserkennung	Das Datum wird erkannt und in ein einheitliches Format gebracht.
Textextraktion	Der Text des Beitrages wird ohne störende HTML-Elemente erfasst. Zitierter Text wird herausgefiltert.
Intelligente Aktualisierung	Der iSquare Spider Newsgruppe untersucht nur die seit dem letzten Durchlauf hinzugekommenen Artikel.
Message	Die Message eines Artikels wird erkannt.
Thread-Erkennung	Aus den Informationen über die Message des Artikels, auf den Bezug genommen wird, kann ein Thread rekonstruiert werden.
Datenbasis	Derzeit kann aus ca. 60.000 relevanten Newsgruppen eine Auswahl getroffen werden.



6. iSquare Spider Email

Der Spider Email beobachtet alle eingehenden Nachrichten einer oder mehrerer Emailkonten.

Feature	Beschreibung
Quellen	Dieser Spider verarbeitet von beliebig vielen Accounts alle ankommenden Mails und deren Attachments.
Attachment- verarbeitung	Attachments in folgenden Formaten können verarbeitet werden: Word, PDF, Excel, Open Office, RTF
HTML-Email	HTML-Emails werden automatisch zur weiteren Verarbeitung in ein Textformat konvertiert.
Attributerkennung	Extraktion von Betreff, Sendedatum und Absender