Suchmaschinen im Kontext von Business Intelligence

Competitive Intelligence

Suchmaschinen werden hauptsächlich mit der Suche in unstrukturierten Datenansammlungen wie Texten assoziiert. Die Eignung für die Analyse strukturierter Informationen ist derzeit dagegen nicht umfassend bekannt. Bei der Verwendung strukturloser Daten stoßen BI-Lösungen leicht an ihre Grenzen, besonders wenn ein eindeutiger Kenner wie beispielsweise ein EAN-Code (Barcode) für eine Preisvergleichsanalyse fehlt.

Competitive Intelligence (kurz CI) steht für Konkurrenz- bzw. Wettbewerbsforschung. Das damit gewonnene Wissen über die Stärken und Schwächen von Wettbewerbern kann mit Hilfe eines systematischen Prozesses – beispielsweise beim Preis- bzw. Sortimentsvergleich – analysiert werden, etwa um strategische Wettbewerbsvorteile zu erzielen. CI nutzt hierfür hauptsächlich unstrukturierte Daten, im Gegensatz zu BI, das vorwiegend unternehmensinterne und strukturierte Informationen verwendet. Damit stellt das Internet für CI eine interessante Informationsquelle dar. In diesem Artikel wird die Internetrecherche, ein Teilaspekt von CI, beleuchtet sowie die Frage, wie sich Suchmaschinen-Technologien für Analysezwecke einsetzen lassen. Als Beispiel dient ein Online-Weinhändler. Dieser möchte seine Preise und das Sortiment mit seinen Wettbewerbern vergleichen.

Es stellt sich die Frage, ob es nicht genügt, die Preise mit Hilfe von Google zu ermitteln. Für den privaten Gebrauch reicht das durchaus. Für einen regelmäßigen, automatisierten Preisvergleich über das gesamte Weinsortiment eines Händlers ist das jedoch nicht die geeignete Methode. In Google finden sich obendrein keine Liefermengen. Auch das Aufspüren von Sortimentslücken oder die Erstellung einer Preishistorie für ein Produkt ist mit Google nicht realisierbar.

Web-Scraping und seine technischen Grenzen

Als Basis für eine effiziente Methode, die in diesem Artikel beleuchtet wird, dient Web-Scraping: ein Verfahren, um Informationen aus Web-Seiten zu extrahieren. Ein Nachteil der Web-Recherche gleich vorab: Web-Seiten ändern sich relativ häufig, aus diesem Grund müssen die Extraktions-Skripte regelmäßig angepasst werden. Auf dem Markt gibt es unterschiedlichste Tools für diese Aufgabe. So kann man beispielsweise das Python Open-Source-Framework Scrapy [Scr15] benutzen. Einen alternativen Ansatz stellt die Datenintegrationssoftware Kapow [Kap15] der Firma Kofax dar. Hier erstellt der Anwender über eine interaktive Oberfläche per Drag & Drop die Robots zum Laden der Informationen (siehe Abbildung 1), mit dem Vorteil, dass später – im Vergleich zu einem Skript – bequemer auf Seitenänderungen reagiert werden kann. Des Weiteren unterstützt Kapow auch die Datentransformation.

Eine Vorbedingung für die Datenerhebung aus dem Internet, die erfüllt sein muss, ist, dass sich diese aus den Web-Seiten überhaupt extrahieren lassen. Die Betreiber von Online-Seiten versuchen das auf verschiedene Art und Weise

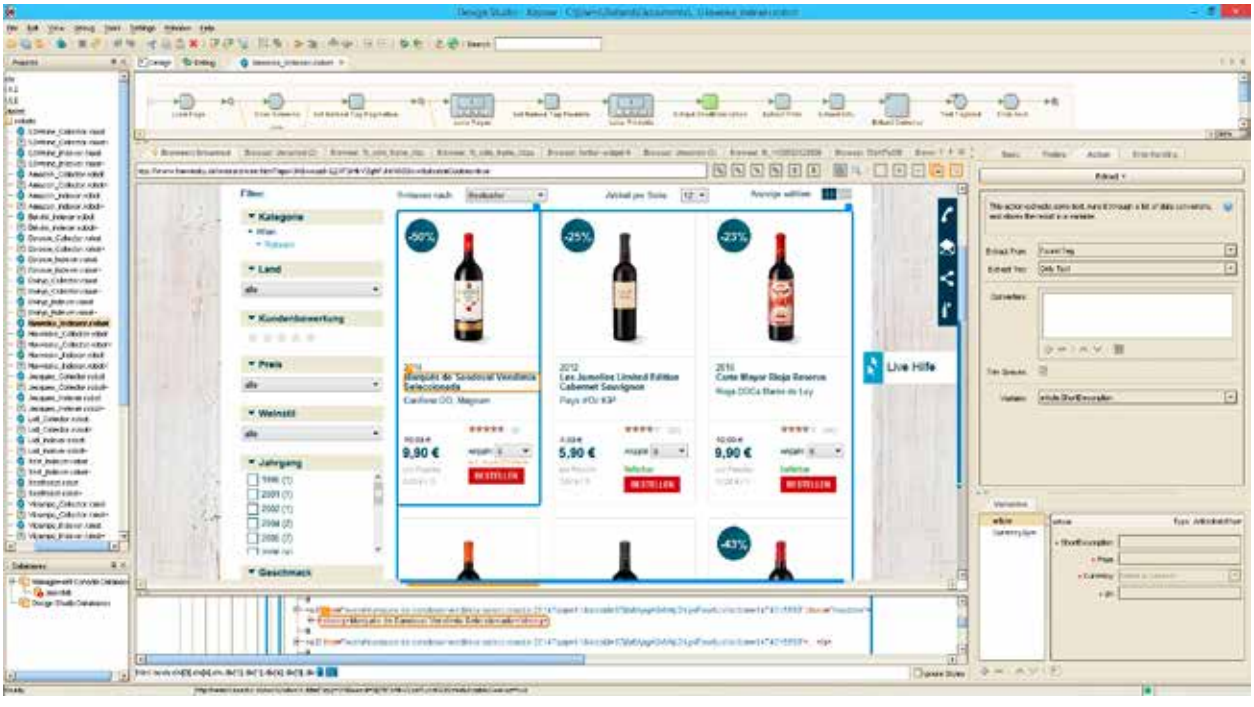

Abb. 1: Erstellung eines Robot mit Hilfe von Kofax Kapow

zu verhindern. Folgende Techniken zum Verschleiern setzt man hierfür gerne ein [Scr13]:

- Zugriff ist durch Login geschützt
- Häufige Änderungen am Seitenaufbau
- Informationen werden nicht als Text, sondern mit Bildern dargestellt
- CAPTCHAS
- Honey Pots – hierbei handelt es sich um versteckte Seiten, die ein Benutzer niemals über das Menü findet, die jedoch von Robots, die das gesamte Verzeichnis des Web-Servers abgreifen, gefunden werden können. Anhand dieser Daten kann man wiederum die Verursacher des Web-Scrapings finden.
- Keine Informationen anzeigen, die nicht abgegriffen werden sollen

In der Praxis zeigt sich, dass die häufige Änderung von Seiten am meisten angewandt wird. Hier helfen intelligent programmierte Agenten zur Extraktion der Inhalte nur begrenzt. Aus diesem Grund ist es bei der Budgetierung eines solchen Vorhabens relevant, die Wartungskosten nicht außer Acht zu lassen.

Extraktions-, Transformations- und Lade-Prozess (ETL)

Die Thematik der Datenqualität macht bei Data-Warehouse-Projekten in der Regel 80 Prozent des Aufwands aus. Dies verhält sich analog zu unserem Beispiel. Die regelmäßige Extraktion der Daten aus den Web-Seiten erfolgt mit dem oben beschriebenen Web-Scraping. Ein wöchentlicher Turnus reicht für einen Preisvergleich im Rahmen einer Wettbewerbsbeobachtung aus. Die meisten Informationen werden direkt aus der Web-Seite extrahiert, manche, wie die Liefermenge, ermittelt man für gewöhnlich über den Einkaufskorb. Dadurch wird der Datenextraktionsprozess komplexer und beansprucht mehr Zeit.

Anschließend werden die gesammelten Daten in ein einheitliches Datenschema transformiert. In diesem Schritt finden die syntaktische (zum Beispiel Datumswerte) und die semantische Wandlung statt. Als Beispiel für den zuletzt genannten Punkt kann die Weinkategorie genannt werden. So wird etwa „Weisswein“ zu „Weißwein“ umbenannt oder die Sterne-Kundenbewertung in Zahlen umgerechnet. Im Anschluss daran werden die Ergebnisse in die Datenbanktabelle geladen. Eine Tabelle ist für unsere Fragestellung vollkommen ausreichend und bietet den Vorteil, dass die spätere Auswertung einfacher wird.

Analyse der Daten

Schließlich kann die Betrachtung der in der Datenbank verfügbaren Seiteninformationen beginnen. So können zum Beispiel Preis- bzw. Weinkategorien der einzelnen Händler verglichen werden, um festzustellen, in welchen Bereichen es sinnvoll ist, das eigene Warenangebot zu erweitern, oder ob ein Mitbewerber bereits Weine des Weinguts XY im Angebot führt. Zum Einsatz kommen hier BI-Werkzeuge wie QlikView oder Cognos.

Leider ist es mit diesen Tools nicht praktikabel, eine weitergehende Analyse wie einen direkten Preisvergleich des Sortiments mit dem der Wettbewerber durchzuführen. Wie eingangs beschrieben, verfügen die meisten Weine über keinen eindeutigen Kenner wie den EAN-Code (Barcode). Demzufolge ist die Benutzung des Artikelnamens nicht zielführend, da jeder Online-Händler diesen unterschiedlich aufbaut. So benutzt der eine beispielsweise den Jahrgang oder den Namen des Weinguts am Ende der Artikelbezeichnung, ein anderer am Anfang. Die bekannten BI-Werkzeuge reichen in diesem Fall nicht aus. Suchmaschinen arbeiten hingegen problemlos mit Texten. Deshalb stellt sich die Frage, ob diese hier helfen und möglicherweise ebenso für Analysen verwendet werden können.

Business Intelligence mit Suchmaschinen

Ein Dokument durchläuft auf dem Weg in den Such-Index einen Analyseprozess. Hierbei wird der Tokenizer involviert, der eine Aufspaltung in einzelne Wörter – beispielsweise durch Leerzeichen getrennt – vornimmt. Anschließend kann mit Hilfe eines Index das Wort wiederum dem zugehörigen Text zugeordnet werden. Folglich ist es für eine Suchmaschine unbedeutend, in welcher Reihenfolge der Jahrgang in dem obigen Beispiel erscheint. Der Like-Operator in SQL ist nicht so flexibel und eignet sich nur, um ähnliche Werte mit Hilfe eines Wildcard-Operators (% bzw. _) zu finden.

In ihrem Buch „Search Based Applications“ [GrW10] zeigen die Autoren Gregory Grefenstette und Laura Wilber auf, wie Suchmaschinen gewaltige Mengen von unstrukturierten und strukturierten Daten nahezu in Echtzeit zur Verfügung stellen können. Dabei verarbeiten sie Web-Seiten und Datenbankinhalte gleichzeitig. Im Open-Source-Umfeld wird die Suchmaschine Elasticsearch [Ela15] immer populärer. Sie verfügt über folgende wesentlichen Funktionen bzw. Einschränkungen:

- Basiert auf dem Open-Source-Projekt Apache Lucene. Es stehen Clusterfunktionen für die rasche und effiziente Verwaltung enormer Datenmengen (Big Data) zur Verfügung. Bei steigender Datenmenge können zusätzliche Knoten hinzugefügt werden.
- Die Persistierung erfolgt wie bei NoSQL-Datenbanken in Form von JSON-Dokumenten.
- Statistische Standard-Analysefunktionen (wie SUM, COUNT, MIN, MAX, AVG, ...) stehen zur Verfügung.
- Die Zuverlässigkeitsgarantien (ACID) sind im Gegensatz zu relationalen Datenbanken (RDB) eingeschränkt.
- Unterstützt die Speicherung und Suche von Geoinformationen bei viel einfacherer Umsetzbarkeit im Vergleich zu namhaften RDBs.
- Im Vergleich zu anderen Big-Data-Tools wie Hadoop zeitsparender aufzusetzen

Folglich ergeben sich verschiedene Einsatzmöglichkeiten, angefangen von einem Lese-Cache für RDBs – insbesondere hierarchische Abfragen erfahren eine Geschwindigkeitsverbesserung – bis hin zu einem Data Warehouse (DWH). Elasticsearch eignet sich demzufolge auch für die Analyse multidimensionaler Informationen. Dabei entfällt die von einem klassischen DWH bekannte nächtliche Voraggregation: Man erhält quasi ein Realtime-DWH.

Bei allen etwaigen Vorteilen gilt es jedoch zu beachten, dass sich die hohe Geschwindigkeit von Elasticsearch – ver-

glichen mit Hadoop – bei der Auswertungsgenauigkeit umgekehrt proportional verhält, allerdings nur bei umfangreichen Datenmengen, die eine Verteilung auf viele Server benötigen.

So ermittelt beispielsweise jeder Knoten für sich die Liste der Top-5-Kunden. Diese Top-5-Klienten stimmen als Einzelergebnis, aber nicht zwingend über alle Knoten des Clusters. Elasticsearch konsolidiert demzufolge nicht die geteilten Analyse-Ergebnisse. Diese Einschränkung gilt nicht, wenn nur ein Knoten im Einsatz ist. Bei Big-Data-Projekten zählt jedoch meistens mehr der Gesamtzusammenhang der Informationen als der einzelne Datensatz. Oder anders formuliert: Mit Big Data wird man vermutlich kein Konzern-Reporting für die Jahresbilanz aufsetzen wollen.

Analyse der Weinhändler mit Hilfe einer Suchmaschine

Für die Auswertung der Daten wird die Datenbank mit Elasticsearch indiziert. Die Abbildung von Dimensionen erfolgt im klassischen DWH unter Zuhilfenahme des Star-Schemas. In Elasticsearch geschieht dies dynamisch über Aggregationen, mit deren Hilfe nach Weinkategorien, Winzern, Ländern, Preisbereichen etc. gefiltert werden kann, analog zum „Drill-down“ und „Roll-up“ in Pivot-Tabellen. Zusätzlich benutzt man die Volltextsuche (inklusive „Meinten Sie“ oder Fuzzy-Suche), um in den Beschreibungstexten der Weine zu suchen (siehe Abbildung 2). Eine grafische Darstellung der Daten ist ebenfalls möglich, und eine Preishistorie wird mit Hilfe der Artikel-URL aufgebaut. Diese Funktionen sind – bis auf die Volltextsuche – ebenso mit klassischen BI-Tools abbildbar. Interessant wird es aber, wenn man Preise ohne einen eindeutigen Kenner wie den EAN-Code (Barcode) vergleichen will.

Um diese Anforderungen abbilden zu können, muss ein Skript entwickelt werden, beispielsweise mit der Sprache Groovy oder Python. Es nimmt den Artikelnamen aus der Datenbank, sucht diesen in der Suchmaschine und fügt jeden Treffer dem neu erstellten Such-Index hinzu. Wie oben erläutert, ist es in diesem Fall egal, ob der Jahrgang am Anfang oder am Ende der Artikelbezeichnung erscheint, da für Elasticsearch die Reihenfolge keine Rolle spielt. Bei der Analyse ist es natürlich unerlässlich, dass man seine Fachdomäne kennt. In unserem Szenario gibt es beispielsweise vom Weingut Rothschild preisgünstige, aber auch teure Weine, die ähnlich heißen.

Das Resultat kann unter Zuhilfenahme der Preisdifferenz verbessert werden. Hierzu wird die Relevanz eines Dokuments im Suchergebnis (bei Suchmaschinen Boosting genannt) erhöht, sofern die Differenz der Preise nicht höher als etwa 30 Prozent ist. Eine abschließende Verifikation der Ergebnisse ergab, dass 80 Prozent der betrachteten Artikel konkret über einen Preisunterschied verfügen. Dies ist ein positives Endergebnis, wenn man bedenkt, dass der Vergleich der Beträge per Text und ohne eindeutige Nummer stattfand. Der gesamte Verarbeitungsprozess ist in Abbildung 3 schematisch dargestellt.

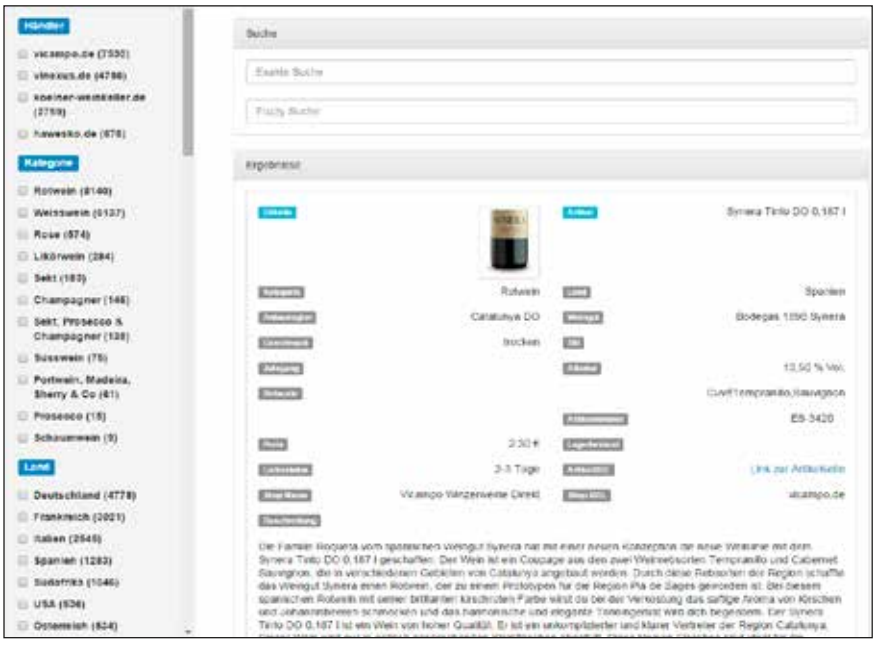

Abb. 2: Datenexploration mit Hilfe der Suchmaschine Elasticsearch

Rechtliche und moralische Aspekte

Im Gegensatz zur Wirtschaftsspionage befasst sich Competitive Intelligence mit Nachdruck nur mit legalen, datenschutzkonformen, öffentlich zugänglichen und ethisch einwandfreien Informationen zu Schwächen, Absichten und Fähigkeiten von Wettbewerbern [WIK15]. Das Amt für Statistik Berlin-Brandenburg [DES13] beschreibt folgende rechtlichen Aspekte:

- Auch die Inhalte von Datenbanken sind urheberrechtlich geschützt, sofern wesentliche Teile betroffen sind.
- Das zielgerichtete Auslesen von Preisen ist unproblematisch.
- AGBs, die Web-Scraping untersagen, sind nur wirksam, wenn sie vorher explizit akzeptiert wurden.
- Es werden nur Daten ausgelesen, die ohne die Akzeptanz von AGBs erreichbar sind.
- Es dürfen keine technischen Hürden umgangen werden (zum Beispiel CAPTCHAS).
- Web-Scraping ist aus diesen Gründen zum Beispiel bei Ryanair nicht umsetzbar.

Ryanair untersagt beispielsweise unmissverständlich das automatische Extrahieren von Informationen in seinen AGBs. Jedoch erließ das BGH am 30. April 2014 ein Urteil, in dem Anbieter von Flugsuchmaschinen das Recht zugesprochen wurde, sich Daten von der Ryanair-Homepage abzugreifen [Zie10; Deu14].

Fazit

Der Einsatzbereich von Suchmaschinen geht weit über die Volltextsuche hinaus. Ihre Vorteile gegenüber klassischen

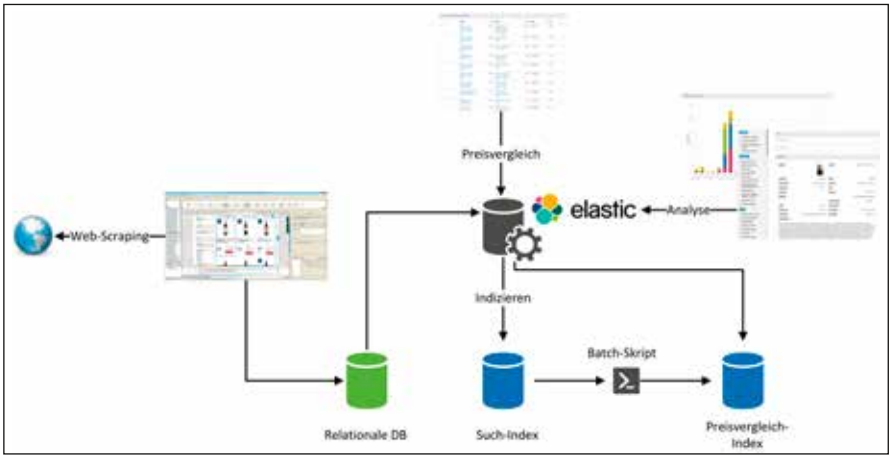

Abb. 3: Gesamter Verarbeitungsprozess

BI-Lösungen können sie besonders effektiv ausspielen, wenn die Analyse von Informationen aus unstrukturierten und strukturierten Daten gefragt ist. Obendrein stehen die Daten nahezu in Echtzeit zur Auswertung zur Verfügung, und dies ohne nächtliche Voraggregationen. Und zu guter Letzt werden zusätzlich Projektkosten reduziert, sofern eine Open-Source-Variante wie Elasticsearch zum Einsatz kommt.

[Literatur / Links]

[DES13] Machbarkeitsstudie zur Automatisierten Preiserhebung im Internet: www.statistik-berlin-brandenburg.de/home/pdf/preise/2013/Bayer_PraesentationInternetpreiserhebungMessungDerPreise.pdf, abgerufen am 28.7.2016

[Deu14] Deutsch, A.: BGH hält „Screen-Scraping" für zulässig – Reiseportale dürfen Ryanair-Flugdaten auslesen. www.lto.de/recht/hintergruende/h/bgh-urteil-izr22412-screen-scraping-flugdaten-automatisiert-auslesen-ryanair-reiseportal/, 30.4.2014, abgerufen am 28.7.2016

[Ela15] Elasticsearch-Suchmaschine: www.elastic.co/de/, abgerufen am 28.7.2016

[GrW10] Grefenstette, G. / Wilber, L.: Search-Based Applications: At the Confluence of Search and Database Technologies. Morgan & Claypool Publishers 2010

[Kap15] Kapow Web Scraping Software. www.kofax.de/datenintegration-extraktion, abgerufen am 28.7.2016

[Scr13] Making Your Website Scrape-Proof. www.speedawarenessmonth.com/making-your-website-scrape-proof-how-to-thwart-content-thieves/, 5.9.2013, abgerufen am 28.7.2016

[Scr15] Scrapy Web Scraping Tool: http://scrapy.org/, abgerufen am 28.7.2016

[WIK15] Wikipedia: Competitive Intelligence, https://de.wikipedia.org/wiki/Competitive_Intelligence#Ethische_Aspekte, abgerufen am 28.7.2016

[Zie10] Ziegler, P.-M.: Landgericht entscheidet im Screen-Scraping-Rechtsstreit: www.heise.de/ct/artikel/Kratzbuerstig-1111030.html, 23.10.2010, abgerufen am 28.7.2016

BI-SPEKTRUM ist eine Fachpublikation des Verlags:
SIGS DATACOM GmbH | Lindlaustraße 2c | 53842 Troisdorf
Tel.: +49 (0) 22 41.2341-100 | Fax: +49 (0) 22 41.2341-199
E-mail: info@sigs-datacom.de
www.javaspektrum.de | www.objektspektrum.de
www.bi-spektrum.de

SIGS DATACOM
FACHINFORMATIONEN FÜR IT-PROFESSIONALS

Peter Soth arbeitete nach seinem Studium der Elektrotechnik und Informatik mehrere Jahre als Software-Engineer und Senior Consultant bei international tätigen Unternehmen wie Hewlett Packard, SAS Institute und BEA Systems (jetzt Oracle). Im Jahre 2006 gründete er die exensio GmbH mit. **E-Mail: peter.soth@exensio.de**