„Match Me If You Can“ – Sammeln und semantisches Aufbereiten von Fußballdaten*

Moritz Finke und Julian Risch

ABSTRACT

Interviews, Spielstatistiken oder Videoaufzeichnungen sind für Fußballfans zwar zahlreich im Internet verfügbar, aber auf viele verschiedene Websites verstreut. „Semantic Media Mining“ verknüpft nun Fußballdaten aus unterschiedlichen Quellen, bereitet sie semantisch auf und führt sie auf einer einzigen Website zusammen. Dadurch dokumentieren und visualisieren wir mehr als 50 Jahre Fußballgeschichte mit über 500 Mannschaften und 40.000 Spielern der Champions League, sowie der 1. und 2. Bundesliga.

FUSSBALLDATEN IM INTERNET

Zu einem populären Themenbereich wie Fußball steht eine Vielzahl verschiedenartiger Datenquellen zur Verfügung. Außer den offiziellen Websites der Ligen und Vereine existieren Internetauftritte von Sportmagazinen wie Kicker mit redaktionell aufbereiteten, aktuellen Inhalten. Twitter und YouTube aus dem Social Web oder Wikipedia als Enzyklopädie bieten nutzergenerierte Daten als weitere Möglichkeit sich zu informieren.

Semantisch verknüpfte Informationen im Internet (Semantic Web) bieten aufgrund des großen verfügbaren Datenumfangs viel Potential für Anwendungen unterschiedlichster Art. Alle oben genannten Plattformen stellen ihre Daten jedoch nur semi- oder unstrukturiert bereit und beleuchten dabei lediglich einzelne Aspekte. Texte, Statistiken, Bilder und Videos sind zudem auf viele verschiedene Websites verstreut. Dadurch muss sich ein Nutzer, seinen Bedürfnissen entsprechend, auf mehreren Seiten einen Überblick verschaffen.

Interessante Statistiken ergeben sich häufig aber erst aus der Kombination der Daten. Zusammenhänge, die sich auf einen großen Zeitraum oder auf Daten aus mehreren Quellen beziehen, bleiben dem Nutzer größtenteils verborgen. Die Frage nach einem Video vom Fußballspiel mit den meisten vergebenen Karten einer Saison oder die Frage nach der Mannschaft, welche am meisten vom Wechsel von der 2- auf die 3-Punkte-Regel profitiert hat, lässt sich beispielsweise nur sehr aufwendig und über Umwege beantworten.

Unsere Arbeit beschäftigt sich mit der Lösung des beschriebenen Problems, indem sie folgenden Ansatz mit Hilfe von Konzepten des Semantic Web verfolgt: Für jede einzelne Quelle werden die Daten aus ihrem bisherigen Kontext extrahiert. Anschließend werden die Daten vereinheitlicht und in einem Gesamtdatenbestand verknüpft. Dieser Datenbestand wird für vielfältige analytische Anfragen und Visualisierungen aufbereitet.

Als Voraussetzung für die Zusammenführung der Daten ergibt sich das Verknüpfen der einzelnen Informationen mit einer umfangreichen und gut strukturierten Datenquelle aus der Linked Open Data Cloud, wie beispielsweise DBpedia oder Freebase. Basierend auf dieser strukturierten Datenquelle können die gesammelten Daten in neuen Zusammenhängen interpretiert und übersichtlich dargestellt werden. Zusätzlich besteht die Möglichkeit den Datenbestand mit Informationen anzureichern, die nicht in einem direkten Zusammenhang zur betrachteten Fußball-Thematik stehen. Beispielswiese lassen historische Aufzeichnungen des Deutschen Wetterdienstes Rückschlüsse auf das Wetter und insbesondere den Zustand des Fußballfeldes bei einer bestimmten Begegnung zu.

DATENMODELL

Standardisierte Informationsbeschreibungsmittel wie das Ressource Description Framework (RDF) und die Web Ontology Language (OWL) bilden eine der technischen Grundlagen des Semantic Web. RDF ist ein vom Wolrd Wide Web Consortium (W3C) herausgegebener Standard zur Beschreibung von Metadaten im Internet[1]. Dabei ordnet sich dieses Framework in

Autoren:
Moritz Exner, Moritz Finke, Julian Risch, Timo Wagner, Tim Zimmermann
Hasso-Plattner-Institut an der Universität Potsdam/IT-Systems Engineering
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam
+49 331 5509 272
julian.risch@hpi.de

*Im Rahmen eines Seminars „Semantic Media Mining“ haben wir, ein fünfköpfiges Studententeam, uns mit dem Sammeln und Aufbereiten von vielfältigen Fußballdaten beschäftigt. Drei von uns setzen derzeit ihr Studium am Hasso-Plattner-Institut fort, die anderen beiden haben nach ihrem Bachelorabschluss bereits angefangen zu arbeiten. Die Extraktion und das semantische Verknüpfen von Informationen aus heterogenen Quellen ist ein fortdauerndes Forschungsgebiet unter anderem am Lehrstuhl „Internet-Technologien und -Systeme“ von HPI-Direktor Prof. Dr. Christoph Meinel.

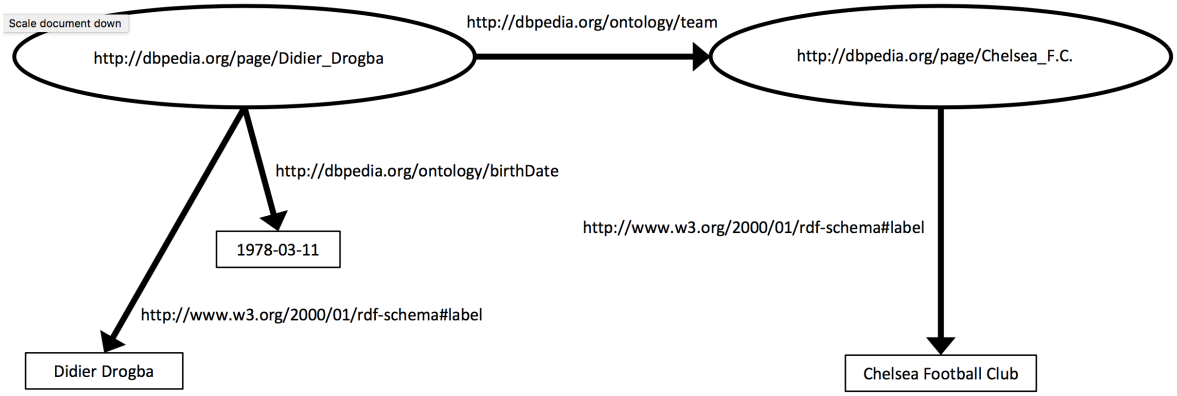

Abbildung 1: RDF Graph mit beispielhaften Fußballdaten

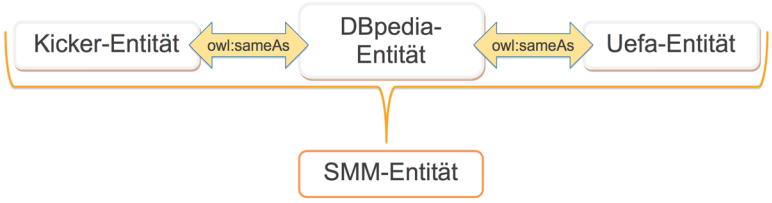

Abbildung 2: Vereinfachte Darstellung der owl:sameAs

das Semantic Web ein, indem es eine Möglichkeit bietet, verteilte, dezentrale Daten zu strukturieren und so zum Beispiel für Web-Anwendungen nutzbar zu machen. Hierbei werden Aussagen in Tripeln der Form <Subjekt, Prädikat, Objekt> getroffen, um beliebige Entitäten als sogenannte Ressourcen zu beschreiben. Das Objekt ist ein einfacher Datenwert (Literal) oder eine Ressource, welche selbst wieder als Subjekt in Erscheinung treten kann. So entstehen Strukturen, die als Graph interpretierbar sind und ein zusammenhängendes System beschreiben (siehe Abbildung 1). Es werden Uniform Resource Identifiers (URIs) nach bestimmten Konventionen verwendet, um eine eindeutige Identifizierung der Ressourcen zu gewährleisten. Ein RDF Schema (RDFS) verwendet RDF zur klassenbasierten Beschreibung eines bestimmten Datenraums. In diesem Zusammenhang wird ein Vokabular für das Beschreiben von Klassen mit Hilfe von RDF definiert.

Die Probleme, die bei der Zusammenführung von heterogenen Daten entstehen, sind detailliert von Bleiholder und Naumann beschrieben (Bleiholder, Jens; Naumann, Felix: Data fusion. In: ACM Comput. Surv. 41 (2009), Januar, Nr. 1, S. 1:1–1:41. – ISSN 0360–0300) und bilden die Grundlage für das Verständnis dieser Arbeit. Zu unterscheiden sind vor allem zwei Probleme: (i) Datenkonflikte, wobei zwei Quellen für dasselbe Attribut eines realen Objektes unterschiedliche Werte vorsehen (Fusion) und (ii) das Problem, festzustellen welche Datensätze unterschiedlicher Quellen dasselbe reale Objekt beschreiben (Matching).

DATENVERWALTUNG

Ein wesentlicher Bestandteil der Arbeit ist das Sammeln und Verknüpfen von Daten aus heterogenen Datenquellen. Dies setzt eine Strategie zur Datenverwaltung voraus. Unsere Arbeit verfolt einen RDF-basierten Ansatz, um die verwendeten Fußballdaten semantisch aufzubereiten und zu verwalten. Wir verwenden bei unserem Projekt das Datenbankmanagementsystem Virtuoso, weil es auf unsere RDF-basierte Anwendung zugeschnitten ist. Es bietet mit einem Quad-Store die Möglichkeit, RDF- Graphen zu benennen und unterstützt mit SPARQL eine mächtige Anfragesprache. Darüber hinaus wartet es aufgrund einer besseren Skalierbarkeit mit Vorteilen im Vergleich zu ähnlichen DBMS-Lösungen auf. Besonders bei großen Datenmengen lassen sich Geschwindigkeitsvorteile feststellen[2].

DATENBASIS

Die Datenbasis der Arbeit ist der englischsprachigen DBpedia[3] entnommen. DBpedia ist ein communitybasiertes Projekt mit dem Ziel, strukturierte Informationen aus Wikipedia zu extrahieren, um sie in einer Virtuoso-Datenbank zur Verfügung zu stellen. Diese Quelle eignet sich für unser Projekt besonders gut, da sie für in Europa aktive Fußballspieler und Vereine Informationen in bereits semantisch aufbereiteter Form anbietet. Daten aus anderen Quellen bilden wir auf die DBpedia als Basis ab. Bei diesem Prozess, der Teil unseres sogenannten Matching-Verfahrens ist, wird einer DBpedia-Entität mittels owl:sameAs-Beziehung eine Entität aus anderen Datenquellen (beispielsweise Kicker Magazin oder UEFA) zugeordnet und umgekehrt (siehe Abbildung 2).

Werden auf diese Weise mehrere Zuordnungen zur selben Entität vorgenommen, so erfolgt die Zuordnung der beteiligten Entitäten durch Bildung der transitiven Hülle. Diese Transitivität wird mit der Same-As-Traversal-Funktion in Virtuoso durch Voranstellen von 'DEFINE input:same-as' in einer SPARQL Anfrage realisiert.

[1] http://www.rdfabout.com/intro/?section=2

[2] http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V6/#comparison

[3] http://dbpedia.org/About

Link	D	EN	ES	IT	EM	WM	F_ D	F_ EM	F_ WM	CL	DFB
de.uefa.com/	-	-	-	-	x	-	-	x	-	x	-
www.kicker.de	x	x	x	x	x	x	x	x	x	x	x
www.youtube.com/user/ SkySportHD	x	-	-	-	-	-	-	-	-	x	x
www.footytube.com/	x	x	x	x	-	-	-	-	-	x	x
de.soccerway.com/	x	x	x	x	x	x	x	x	x	x	x
pesstatsdatabase.com	x	x	x	x	-	-	-	-	-	-	-
twitter.com/Kicker_live	x	-	-	-	x	x	-	-	-	x	x
www.fifa.com/	-	-	-	-	-	x	-	-	x	-	-
www.goalsarena.org/	x	x	x	x	x	x	-	-	-	x	x
www.openligadb.de/	x	x	x	x	x	x	-	-	x	x	x
www.soccerstats.com/	x	x	x	x	x	x	-	-	-	x	x
www.worldfootball.net/	x	x	x	x	x	x	x	x	x	x	x

Link	API	Lizenz	Umfang
de.uefa.com/	-	private Nutzung erlaubt	CL komplett, national nur aktuelle Saison
www.kicker.de/	-	private Nutzung erlaubt	D komplett, andere ab Ende 1990er-Jahre
www.youtube.com/user/ SkySportHD	x	Videoeinbindung erlaubt	Videos zur aktuellen Saison
www.footytube.com/	-	Videoeinbindung erlaubt	teilweise auch Videos vergangener Saisons
de.soccerway.com/	-	private Nutzung erlaubt	BL komplett, CL ab 2000
pesstatsdatabase.com	-	frei	nur Spielerdaten, keine Ergebnisse etc.
twitter.com/Kicker_live	*	frei	Live-Ergebnisse
www.fifa.com/	-	private Nutzung erlaubt	komplett
www.goalsarena.org/	-	private Nutzung erlaubt	ab 1999
www.openligadb.de/	x	frei	ab 2008, aber unvollständig
www.soccerstats.com/	-	private Nutzung erlaubt	ab Mitte 2000
www.worldfootball.net/	-	private Nutzung erlaubt	Frauenfußball ab Mitte der 1990er-Jahre

Tabelle 1: Verschiedene Datenquellen und der Umfang der zur Verfügung stehenden Daten. Beziehung.

Neben DBpedia verwenden wir auch andere Quellen. Eine Übersicht der ursprünglichen Kandidaten, von denen die ersten vier ausgewählt wurden, ist in Tabelle 1 dargestellt.

Die Quellen werden mit niedriger Bandbreite und Pausen zwischen den einzelnen Seitenaufrufen indiziert, um die Last auf der Quelle möglichst gering zu halten. So erhaltene Daten werden in das RDF-Format transformiert, indem HTML-Inhalte extrahiert und mit zusätzlichen Informationen aus anderen Dateien dieser Quelle angereichert werden. So können beispielsweise Fußballspieler mit passenden Informationen wie vollem Namen, Geburtsdatum und Bildern erweitert werden.

Da die Struktur der Daten für verschiedene Zeiträume unterschiedlich sein kann, werden zur Inhaltsextraktion möglichst allgemeine reguläre Ausdrücke verwendet. Ebenso kann sich die Zeichenkodierung der Seiten ändern, was die Umwandlung in einen einzigen Zeichensatz notwendig macht.

Zusätzlich zur Zuordnung offensichtlicher Fakten bezüglich einer Entität, können auch erste semantische Verknüpfungen in der Transformation stattfinden. Beispielsweise war bei einem Spiel des FC Chelsea am 19.05.2012 Didier Drogba in der Startaufstellung. Diese Information wird bei unserem Projekt unter anderem auf ein Tripel mit <Didier Drogba, club-2012, FC Chelsea> gebildet. Von der Startaufstellung einer Mannschaft zu einem bestimmten Datum extrahiert unser Ansatz also auch die Zugehörigkeit der einzelnen Spieler zur Mannschaft. Dies erleichtert später die Berechnung einer Spielerlaufbahn.

Des Weiteren können Überschriften oder wiederkehrende Passagen, beispielsweise Stimmen zum Spiel, aus Spielberichten extrahiert und gespeichert werden. Insbesondere Überschriften fassen meist spielentscheidende Aktionen oder ganze Spiele kurz und prägnant zusammen.

Eine essentielle Aufgabe für die Zusammenführung und semantische Aufbereitung der gesammelten Daten ist das sogenannte Entity-Matching. Diese Herausforderung lässt sich als formales Problem spezifizieren. Gegeben seien zwei Datenquellen S_a und S_b und die zugehörigen Entitätsmengen $A \in S_a$ sowie $B \in S_b$. Dann sollen die Entitäten aus A und B, welche jeweils dasselbe reale Objekt beschreiben, einander zugeordnet werden. Das Matching setzt alle Entitäten in eine Äquivalenzrelation. Die damit verbundenen Transitivitäts- und Symmetrieeigenschaften sind Voraussetzungen dafür, dass das Problem auf das Abbilden in einen zentralen Basisdatensatz zurückgeführt werden kann.

Da die von uns verwendeten Quellen im Wesentlichen duplikatfrei sind, ist es nicht nötig, Entity-Matching innerhalb einer Datenquelle zu betreiben. Wir gehen davon aus, dass zwei beliebige Entitäten innerhalb einer Quelle stets unterschiedliche reale Objekte beschreiben. Die Schwierigkeit besteht vielmehr darin, Entitäten aus heterogenen Datenquellen eindeutig und vor allem korrekt im Sinne der oben definierten Vorgaben einander zuzuordnen.

Für eine eindeutige Zuordnung mit Hilfe der vorliegenden Attribute (beispielsweise Spielername) spielen unter Umständen Kriterien zur Einschränkung des Ergebnisraums eine Rolle. Für den Fall, dass kein Geburtsdatum eingetragen ist, kann kein eindeutiges Matching garantiert werden. Beispielsweise existiert in unserem Basisdatensatz ein Eintrag zu einem Spieler, der den vollen Namen des Spielers und sein Geburtsdatum enthält. Wenn bei einem Fußballspiel ein Spieler mit diesem Namen erwähnt wird so kann zusätzlich zur Namensübereinstimmung überprüft werden, ob der Spieler zum Zeitpunkt des gegebenen Spiels in einem spielfähigen Alter war oder ob er zu diesem Zeitpunkt noch gar nicht geboren war.

Die Qualität der zusammenzuführenden Datenquellen ist ein wesentlicher Faktor für ein erfolgreiches Matching. Speziell die Hierarchie und Beschaffenheit der DBpedia-Datensätze sind im Zusammenhang dieser Arbeit interessant. Im Folgenden wird dargestellt, wie diese Eigenschaften eingesetzt werden, um typischen Schwierigkeiten beim Matching im Anwendungsbereich Fußball entgegenzuwirken.

Das Kernproblem stellen unterschiedliche Bezeichnungen für die gleiche, reale Entität dar. So ist zum Beispiel in manchen Datenquellen von ”FC Chelsea“ und in anderen von ”The Blues“ die Rede, wenn der Champions League-Sieger der Saison 2011/12 gemeint ist. Abweichungen in der Benennung können aber nicht nur durch Spitznamen, sondern auch durch unterschiedliche Schreibweisen entstehen.

In der Ontologie von DBpedia existieren viele unterschiedliche Entitätstypen, die einem Objekt einen oder mehrere Namen zuordnen. Um Namenskonflikte aufzulösen, verwenden wir Attribute wie ”rdfs:label“, ”dbprop:name“ und ”dbo:wikiPageRedirects“, wobei jeder dieser Entitätstypen gleichberechtigt in das Matching einbezogen wird. Damit findet im ersten Schritt eine Art Ontology-Matching zwischen den namensgebenden Attributen statt.

Darüber hinaus gibt es in der DBpedia unterschiedliche Möglichkeiten, Entitäten in einen speziellen Kontext einzuordnen. In unserer Arbeit optimieren wir das Matching mit Rücksicht auf die zuvor genannten Ansätze. Dazu werden die namensgebenden Attribute aus der DBpedia mit einem regulären Ausdruck abgeglichen, der auf dem Namen des jeweiligen Objekts basiert. Wird beispielsweise der FC Chelsea in den ”dbo:wikiPageRedirects“ auch als ”The Blues“ geführt, so kann die Mannschaft auch über diesen Namen zugeordnet werden. Zusätzlich wird der betrachtete Entitätsbereich beispielsweise durch Listen oder Kategorien wie ”dbo:SoccerClub“ und ”dbo:SoccerPlayer“ eingeschränkt. Der anschließend verwendete reguläre Ausdruck darf zu Gunsten einer erhöhten Matching-Rate durchlässiger werden, ohne dass dabei die Gefahr von falsch oder mehrfach zugeordneten Entitäten erhöht wird.

MATCHING

Beispiel des Mannschafts-Matching

Für das Team-Matching werden zuerst sämtliche Nicht-ASCII-Zeichen innerhalb des jeweiligen Mannschaftsnamens durch einen Punkt ersetzt. Dadurch entsteht ein regulärer Ausdruck, der für Anfragen an DBpedia verwendet wird. Diese Anfrage besteht neben dem Abgleich der Namen auch aus einer Spezifizierung des Entitätstyps, nämlich dbo:SoccerClub, und der Überprüfung, ob es sich auch wirklich um einen Fußballverein handelt. Dies ist notwendig, da neben Fußballmannschaften fälschlicherweise auch viele andere Sportvereine in dieser Kategorie in der DBpedia vertreten sind. Wir überprüfen, welcher Liga ein Verein zugeordnet ist. Konkret muss die Liga ”Template:Infobox football league“ verwendet werden.

Zusätzlich können hier auch noch spezielle Textkürzel implizit ausgeschlossen werden. Sinnvoll ist dies vor allem zum Ausschluss von Reserve- oder falsch kategorisierten Mannschaften einer anderen Sportart.

Findet diese Anfrage genau einen passenden Verein, so wird dieser in die Datenbank geladen und mittels sameAs mit dem angefragten Objekt beidseitig verknüpft.

Bei mehreren Treffern werden Teilwörter mit drei oder weniger Buchstaben eliminiert und durch .* beliebige Wörter zwischen den resultierenden Teilwörtern erlaubt. Dies hat den Grund, dass die verwendeten Mannschaftskürzel wie FC oder VfB an unterschiedlichen Plätzen innerhalb des Namens stehen können.

Bei keiner oder mehreren gefundenen Mannschaften wird eine zusätzliche Anfrage gestellt. Diese ist weitestgehend analog zur ersten Anfrage, allerdings wird auf eine Überprüfung des verwendeten Ligentemplates verzichtet. Gibt sie mehrere Teamnamen zurück, so wird der kürzeste Name als Treffer betrachtet. Dieser Treffer wird dann geladen und mit sameAs bidirektional verknüpft. Die Betrachtung von Teams mit dem kürzesten Namen hat sich als notwendig erwiesen, da es einige Vereine gibt, die den selben Namen plus eine Staats- oder Stadtbe-

Abbildung 3: Screenshot der entstandenen Website nach der Zusammenführung verschiedener Datenquellen.

schreibung haben. Ein Beispiel hierfür ist der ”Chelsea F.C.“ . Neben diesem Verein kann ebenfalls ”Berekum Chelsea F.C.“ gefunden werden. Dies hat vor allem bei der Champions League Bedeutung, wohingegen bei den deutschen Ligen über Kategorien wie ”Category:Football clubs in Germany “ oder ”GermanFootballCluBs“ in YAGO der Wertebereich eingeschränkt werden kann.

In selten Fällen, in denen für eine Mannschaft kein Matching durchgeführt werden kann, findet eine manuelle Zuordnung statt. Dies ist allerdings nur der Fall, wenn der Teamname stark von den entsprechenden Namen in der DBpedia abweicht. Betroffen sind unter anderem Mannschaften, die in den letzten Jahrzehnten neugegründet oder mit anderen Vereinen zusammengeschlossen wurden. Ein Beispiel hierfür ist der Club ”Vorwärts Leipzig“, der später ”Vorwärts Berlin“, bis 2012 ”Frankfurter FC Victoria“ hieß und sich jetzt mit dem ”MSV Eintracht Frankfurt“ zum ”1. FC Frankfurt E/V“ zusammengeschlossen hat.

Beispiel des Spieler-Matching

Als Beispiel für das Spieler-Matching wird ein Spieler namens Sergij Dkhtjar mit dem Ge- burtsdatum 26.08.1975 verwendet. Die Anfrage nach Namen der Fußballspieler mit diesem Geburtsdatum ergibt folgendes Ergebnis (Gruppiert in Entitäten):

- ’Âîñêàíÿí, Àðàì Ãàéêîâè÷’, ’Aram Voskanyan’
- ’Akide’, ’Mercy Akide-Udoh’, ’Mercy Akide’, ’Mercy Akide Udoh’
- ’Momar Njie’
- ’Sergei Dichtjar’, ’Sergej Dichtiar’, ’Sergei Dikhtyar’, ’Sergej Dikhtjar’, ’Sergey Dikhtyar’, ’Serhij Dychtjar’, ’Serhij Dychtjar’, ’Serhij Dichtiar’, ’Serhij Dychtjar’, ’Serhiy Dikhtiar’
- ’Timur Yanyali’

Anschließend wird durch Substitution der Buchstaben ”aeiouyjncsz “ mit einer beliebigen Zeichenkette (.*) ein regulärer Ausdruck erstellt. Zusätzlich werden beliebige Wörter vor, zwischen und nach den Namensteilen erlaubt. Da dieser Spielername keine Mittelnamen oder doppelten Buchstaben enthält, müssen keine weiteren Vorkehrungen getroffen werden. Das Resultat sieht also folgendermaßen aus: Sergij Dkhtjar -> .*S.*rg.*D.*kht.*r.* Nun wird dieser reguläre Ausdruck auf die Treffer der Anfrage an DBpedia angewendet und diese nach ihrer Levenshtein-Distanz zum ursprünglichen Namen sortiert:

- ’Sergej Dikhtjar’ (Levenshtein-Distanz 0)
- ’Sergei Dikhtyar’ (Levenshtein-Distanz 2)
- ’Sergey Dikhtyar’ (Levenshtein-Distanz 2)

Das gesuchte Objekt wird auf die Entität des Eintrags mit der kleinsten Distanz abgebildet.

WEBSITE

Der Fokus der Webseite liegt auf der Verknüpfung der Daten mit multimedialen Inhalten wie Videos und Diagrammen. Auf den Seiten zu einem Spiel werden beispielsweise Informationen über vorherige Begegnungen der beiden Mannschaften angezeigt und ausgewertet, sodass sich Nutzer leicht einen Überblick verschaffen können. Die Website ist beispielhaft in Abbildung 3 dargestellt. Darüber hinaus wird das jeweilige Spiel in seinen zeitlichen Zusammenhang eingeordnet, um die zum Spieltag aktuelle Tabelle und andere Partien anzuzeigen. Zusätzlich bilden multimediale Inhalte, wie Videos, einen wesentlichen Bestandteil der dargestellten Inhalte. Die Kombination dieser Elemente fehlt auf bestehenden Seiten und ist ausschließlich in unserem System umgesetzt. Bestehende Seiten beschränken sich meist

auf Teilaspekte, sodass Nutzer auf Informationen mehrerer Seiten zurückgreifen müssen, um sich einen Gesamtüberblick zu verschaffen. Unsere Arbeit informiert auf einer einzigen Seite und möglichst facettenreich über die betrachteten Entitäten, um neue, individuelle Perspektiven auf die Daten zu eröffnen. Darüber hinaus besteht für einen Nutzer die Möglichkeit, in Diagrammen die für ihn persönlich interessanten Statistiken einzublenden. Auf den Seiten zu einer kompletten Saison ist dies in besonderem Maße möglich. Hier dienen die Tabellen von jedem Spieltag als Datenquelle für ein Diagramm. Besonders zeitliche Veränderungen über den Saisonverlauf lassen sich so graphisch beobachten. Auf existierenden Fußball-Websites fehlt dieses Maß an Individualität und Benutzerfreundlichkeit oder ist weniger ausgeprägt.

Beispielhafte Ergebnisse

Durch die historische Vollständigkeit der Daten können wir Statistiken abrufen, die ansonsten nur sehr schwer zu generieren sind. So lässt sich beispielsweise ermitteln, dass in der Zeit zwischen der ersten Bundesliga-Saison 1963/64 und der Saison 2011/12 Matthias Scherz die meisten Tore nach der Einwechslung erzielte (insgesamt 19), Jürgen Kreyer mit 6,23 Karten pro Spiel der strengste Schiedsrichter war und Borussia Dortmund in der Saison 1995/1996 die meisten Tore in der ersten Halbzeit erzielte (insgesamt 32). Die besondere Schwierigkeit liegt darin, dass die Bezeichnungen sehr verschieden sein können. Unsere Software muss zum Beispiel erkennen, dass Christiano Ronaldo häufig mit CR7 abgekürzt wird und sich die Informationen auf ein und dieselbe Person beziehen.

Fritz-Walter-Wetter

Fritz Walter war ein deutscher Fußballspieler, der unter anderem bei der WM 1954 als Kapitän der Nationalmannschaft antrat. Er bevorzugte regnerisches Wetter bei Fußballspielen und war insbesondere auf nassem Boden seinen Gegnern überlegen. Fritz-Walter-Wetter bezeichnet demzufolge regnerisches Wetter in Anlehnung an das Finale der WM 1954 . Unser System kann durch die Einbindung von Wetterdaten des Deutschen Wetterdienstes auch folgende Frage beantworten: Welches war das torreichste Spiel, bei dem es geregnet hat? Durch eine weitere Verknüpfung der Datenquellen liefert unser System darüberhinaus gleich ein Video des Spiels.

ZUSAMMENFASSUNG UND AUSBLICK

Wir zeigen, wie viele unterschiedliche mediale Inhalte zusammengeführt und auf einer Website zusammen mit passenden Tweets und Wetterdaten kombiniert zur Verfügung gestellt werden können. Der Schwerpunkt liegt darauf, Informationen aus verschiedenen Quellen über ein und dasselbe Objekt semantisch aufzubereiten und zu verknüpfen (Matching). Ein Objekt kann dabei beispielsweise eine Fußballmannschaft oder ein Fußballspieler sein. Unser Ansatz vereinheitlicht unterschiedliche Bezeichnungen für dasselbe Objekt durch die Betrachtung von Name und Geburtsdatum bei Fußballspielern als Schlüsselattribute. Wir stellen Verfahren vor, die diesen Ansatz unter Verwendung der Levenshtein-Distanz und Algorithmen zur Normalisierung der Objektbezeichner realisieren. Damit lassen sich die verteilten Informationen zusammenführen, als Gesamtheit analysieren und übersichtlich präsentieren, um interessante und detaillierte Aussagen über die letzten 50 Jahre Fußballgeschichte treffen zu können. Unser Datenbestand umfasst 575 Mannschaften, 21.000 Spiele und 40.000 Spieler aus Champions League, sowie 1. und 2. Bundesliga. Diese Daten sind durch insgesamt mehr als 190.000 Entitäten und 3,5 Millionen Tripel repräsentiert.

In Zukunft könnten Geodaten (zusätzlich zu den von uns betrachteten Daten) im Zusammenhang mit Mannschaften und Spielern erfasst und analysiert werden. Anhand von Geodaten kann ermittelt werden, in welchen Städten oder Ländern bestimmte Spieler oder Teams besonders erfolgreich sind. Der gleiche Ansatz ließe sich auch mit den Wetterdaten verfolgen. Welche Mannschaften schneiden bei Regen am besten ab oder welcher Spieler ist auch bei niedrigen Temperaturen in Topform? Neben zusätzlichen Inhalten ließen sich durch die Berücksichtigung

weiterer Wettbewerbe, wie der Frauen-Bundesliga oder Weltmeisterschaften, nationale oder geschlechtsspezifische Unterschiede auswerten. Besonders beim Frauenfußball besteht die Möglichkeit, viele Statistiken erstmalig zu erstellen, da der Informationsumfang dort bisher vergleichsweise gering ist.

Die DBpedia-Anfragen könnten mit Hilfe von YAGO oder weiteren Kategorien noch um Spieler ergänzt werden, die bisher durch falsche Kategorisierung innerhalb der DBpedia nicht berücksichtigt wurden. Lokalisierte Varianten der DBpedia könnten darüber hinaus zur Gewinnung eines breiteren Spektrums historischer Daten verwendet werden.

Hinsichtlich der Website könnte der Aspekt von Live-Informationen stärker berücksichtigt werden. Hierfür ließen sich beispielsweise für bereits angekündigte Spiele die entsprechenden Entitäten anlegen, um diese dann mit Echtzeitdaten aus Twitter oder Live-Tickern zu füllen. Zusätzlich könnten Ansätze zur Analyse der Konnotation von Tweets benutzt und verbessert werden, sodass es zum Beispiel ermöglicht wird, einem Spieler, Team oder Spiel Beliebtheitswerte zuzuordnen. Abschließend stellen wir fest, dass mit unserer Arbeit zum Thema Fußball eine solide Basis im Hinblick auf die genannten Erweiterungsmöglichkeiten geschaffen wurde, die wichtige Konzepte des Semantic Web aufgreift und dieses mit einem weiteren Beitrag bereichert. Leider erlauben es viele der von uns genutzten Quellen aus urheberrechtlichen Gründen nicht, dass die so gewonnenen Informationen auch veröffentlicht werden. •

LITERATUR

[BN09] Bleiholder, Jens ; Naumann, Felix: Data fusion. In: ACM Comput. Surv. 41 (2009), Januar, Nr. 1, S. 1:1–1:41. – ISSN 0360–0300

[BLHL01] Berners-Lee, Tim ; Hendler, James ; Lassila, Ora: The Semantic Web. In: Scientific American 284 (2001), Mai, Nr. 5, S. 34–43

[LS11] Lanagan, James ; Smeaton, Alan F.: Using Twitter to Detect and Tag Important Events in Live Sports. In: Artificial Intelligence (2011), S. 542– 545

[OED12] Oorschot, Guido van ; Erp, Marieke van ; Dijkshoorn, Chris: Automatic Extraction of Soccer Game Events from Twitter. In: Erp, Marieke van (Hrsg.) ; Hollink, Laura (Hrsg.) ; Hage, Willem R. (Hrsg.) ; Troncy, Raphaël (Hrsg.) ; Shamma, David A. (Hrsg.): Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012) Bd. 902. Boston, USA : CEUR, 11 2012, S. 21–30

[PS09] Patman, F. ; Shaefer, L.: Is Soundex Good Enough for You? On the Hidden Risks of Soundex-Based Name Searching. 41 (2009), Januar, Nr. 1

[QLW+10] Qian, Xueming ; Liu, Guizhong ; Wang, Huan ; Li, Zhi ; Wang, Zhe: Soccer video event detection by fusing middle level visual semantics of an event clip. In: Proceedings of the Advances in multimedia information processing, and 11th Pacific Rim conference on Multimedia: Part II. Berlin, Heidelberg : Springer-Verlag, 2010 (PCM'10), S. 439–451

ZZWV11] Zhao, Siqi ; Zhong, Lin ; Wickramasuriya, Jehan ; Vasudevan, Venu: Analyzing twitter for social tv: Sentiment extraction for sports. In: Proceedings of the 2nd International Workshop on Future of Television, 2011

*Vortragsmanuskript (gehalten auf der Frühjahrstagung des vfm am 25. April 2017)