Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia

Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, Gerhard Weikum
Max-Planck-Institut für Informatik
Campus E1 4
Saarbrücken, Germany
{ywang, mzhu, lqu, mspaniol, weikum}@mpi-inf.mpg.de

ABSTRACT

Recent progress in information extraction has shown how to automatically build large ontologies from high-quality sources like Wikipedia. But knowledge evolves over time; facts have associated validity intervals. Therefore, ontologies should include time as a first-class dimension. In this paper, we introduce Timely YAGO, which extends our previously built knowledge base YAGO with temporal aspects. This prototype system extracts temporal facts from Wikipedia infoboxes, categories, and lists in articles, and integrates these into the Timely YAGO knowledge base. We also support querying temporal facts, by temporal predicates in a SPARQL-style language. Visualization of query results is provided in order to better understand of the dynamic nature of knowledge.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]

General Terms: Knowledge Management, Ontology

Keywords: Knowledge Harvesting, Temporal Fact Extraction, Wikipedia, Temporal Queries

1. INTRODUCTION

In order to effectively exploit the enormous knowledge available in high-quality Internet sources such as Wikipedia, we need to extract and structure information and develop efficient methods for querying and exploration. For this purpose, we built YAGO [10], a large-scale and highly accurate knowledge base of with convenient search capabilities. We harvested facts about individual entities from infoboxes and categories of Wikipedia articles and unified them with WordNet. Currently YAGO contains about 3 million entities and more than 20 million facts (instances of binary relations between entities). The knowledge is organized in the form of RDF subject-property-object triples, and can be search by a SPARQL-like query language.

So far, facts in YAGO are static: interpreted as if they were time-invariant. However, the world is dynamic. New facts arise, while some facts change over time. For example, Barrack Obama succeeded George W. Bush as US president in 2009. It is necessary to extend the facts in YAGO with temporal information to build an enhanced time-aware ontology: *Timely YAGO* or *T-YAGO* for short. In addition, presenting facts along the timeline can guide users towards better understanding of knowledge evolution.

As possible application scenarios, we intend to support journalists in need of background information about entertainment or sports stars, or politologists and media analysts in their research on relationships and trends in society and politics. For example, *T-YAGO* should know about the biographical facts of soccer stars like David Beckham, his teams, his team mates, his successes like cups and championships, but also his failures like losing a final or missing a penalty, his residences, romantic affairs, marriage, births of children, etc. Analogously, in politics *T-YAGO* should be aware about facts e.g. Hillary Clinton's education, awards and honors, political positions that she held during specific periods, relationships with other politicians, business people, companies, media, etc. Even trips to important political meetings should be incorporated. Queries could ask for her relationships to other people and organizations of interest.

Efficiently extracting temporal facts from arbitrary natural-language texts with high precision is extremely difficult if feasible at all. Therefore, we start by harvesting temporal facts from Wikipedia infoboxes, the category system, and lists in articles. These facts serve as the backbone of our temporal ontology and will be used for bootstrapping the extraction of temporal facts from free text in our ongoing project. To the best of our knowledge, there is no other work on automatically constructing ontologies with specific consideration of temporal facts.

2. OUR SYSTEM

In this section, we present the architecture of our system. We first introduce the data model in *T-YAGO*, then we describe how we gather temporal facts, and we show how we support queries on temporal facts.

2.1 Data Model

A widely used formalism in knowledge representation is OWL. However, it is computationally expensive; instead we use a simpler RDF-style model.

As in OWL and RDF, all objects are represented as *entities* in the YAGO data model. Two entities can stand in a *relation,* and we call this relational instance a *fact*. All facts are represented by unary or binary relations. Unary relations capture memberships in

EDBT 2010, March 22–26, 2010, Lausanne, Switzerland.

semantic types such as: DavidBeckham *instanceOf* SoccerPlayer. Binary relations like *bornOn*, *bornIn*, *marriedTo*, or *hasWonPrize* hold between entities of specific types; an example is:

DavidBeckham *hasWonPrize* UEFAClubPlayerOfTheYear

This fairly simple model has proven to be very valuable for knowledge exploration and querying. However, facts actually have associated time information. For example, "David Beckham has won the UEFA Club Player of the Year" in "1999". Therefore, in *T-YAGO*, we introduce the concept of *temporal facts*. A temporal fact is a relation with an associated *validity time*. The fact may be valid at a time point or within a time interval.

In the current YAGO ontology, temporal facts cannot be directly represented. Facts are limited to binary relations while temporal facts have more than two arguments. To support temporal facts in a binary relation model, we decompose the n-ary fact into a primary fact and several associated facts. We assign a fact identifier to the primary fact, and then the associated facts are represented as the relation between the identifier and the remaining arguments. For example, for the ternary relation between DavidBeckham, the UEFAClubPlayeroftheYear, and 1999, we use the original time-agnostic fact as a primary fact with identifier #1. For temporal facts valid at a time point, we use the relation *on* to describe the validity time. Then we represent the temporal property of the primary fact as: #1 *on* 1999.

For temporal facts that are valid during a time period, we use two relations to represent the interval: the relation *since* for the begin time point, and the relation *until* for the end time point. For example, the fact that Beckham played for Real Madrid from 2003 to 2007 is represented as:

#2: DavidBeckham *playsFor* RealMadrid
#2 *since* 2003
#2 *until* 2007

Sometimes it is impossible to extract accurate timepoints, say the exact day, and sometimes we may know only the begin or the end of a fact's validity interval but not both. For these situations, we adopt the time expression in TOB [13] for timepoints with the *earliest* and *latest* possible time to constrain the range of the true time point. For example, if we know that Beckham started playing for Real Madrid in July 2003 and terminated his contract in 2007, we would add the temporal facts:

#2 *since* [1-July-2003, 31-July-2003]
#2 *until* [1-January-2007, 31-December-2007]

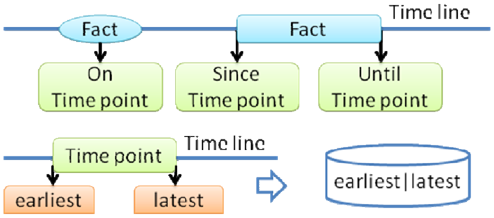


Figure 1. Representation of temporal facts.

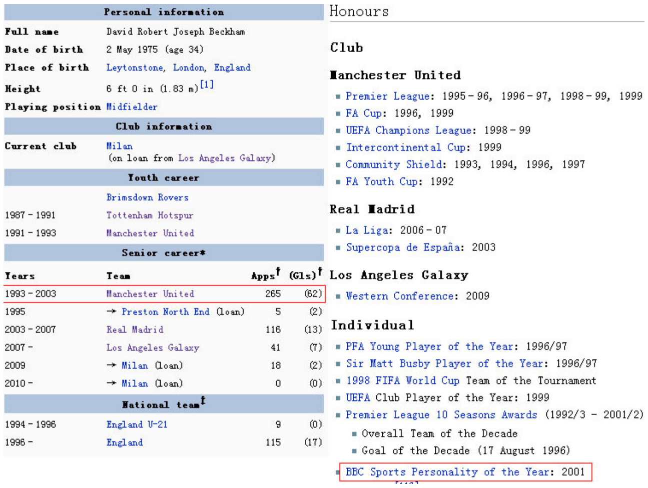
If we later learn that his contract with the team Los Angeles Galaxy started on July 1, 2007, and assuming a constraint that one must not play for two clubs at the same time, we could refine the *until* relation for #2 into:

#2 *until* [1-January-2007, 30-June-2007]

Figure 1 summarizes our representation of validity times of facts.

2.2 Temporal Fact Extraction

Currently YAGO is built upon facts extracted from semistructured infoboxes and category information in Wikipedia, and it is unified with the taxonomic class system of WordNet. Infoboxes are based on templates which are re-used for important types of entities such as countries, companies, sportsmen, politicians, etc. This allows YAGO to apply rule-based extraction for popular infobox attributes (the attributes in the "long tail" are typically skipped as they are highly diverse and noisy). Like YAGO, our approach to temporal fact extraction also focuses on the semistructured elements in Wikipedia. However, as we need to detect also the validity time of each fact, our rules are more sophisticated than those of the original YAGO. We use regular expression matching for this purpose. For higher coverage, we also analyze additional elements in Wikipedia articles, most notably lists such as awards which contain many meaningful temporal facts. Here again, regular expressions tailored to specific types of lists yield high return. In addition, our rule-based techniques can also bootstrap learning-based methods applied on the natural-language text of the full articles. For the latter, the lack of manually labeled training data is often the bottleneck if not a showstopper. Our rich collection of temporal facts in *T-YAGO* can serve as training data and reduce the need for expensive hand-labeling.



Personal information			
Full name	David Robert Joseph Beckham		
Date of birth	2 May 1975 (age 34)		
Place of birth	Leytonstone, London, England		
Height	6 ft 0 in (1.83 m)[1]		
Playing position	Midfielder		
Club information			
Current club	Milan (on loan from Los Angeles Galaxy)		
Youth career			
	Brimsdown Rovers		
1987 - 1991	Tottenham Hotspur		
1991 - 1993	Manchester United		
Senior career*			
Years	Team	Apps†	(Gls)†
1993 - 2003	Manchester United	265	(62)
1995	→ Preston North End (loan)	5	(2)
2003 - 2007	Real Madrid	116	(13)
2007 -	Los Angeles Galaxy	41	(7)
2009	→ Milan (loan)	18	(2)
2010 -	→ Milan (loan)	0	(0)
National team‡			
1994 - 1996	England U-21	9	(0)
1996 -	England	115	(17)

Honours

Club

Manchester United

- Premier League: 1995 - 96, 1996 - 97, 1998 - 99, 1999
- FA Cup: 1996, 1999
- UEFA Champions League: 1998 - 99
- Intercontinental Cup: 1999
- Community Shield: 1993, 1994, 1996, 1997
- FA Youth Cup: 1992

Real Madrid

- La Liga: 2006 - 07
- Supercopa de España: 2003

Los Angeles Galaxy

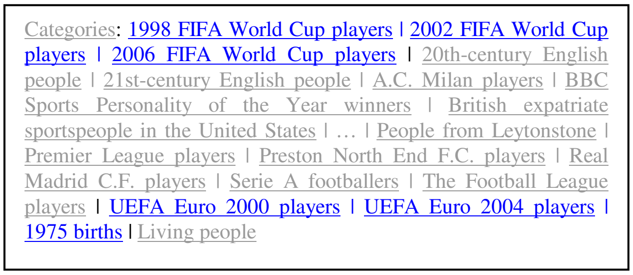
- Western Conference: 2009

Individual

- PFA Young Player of the Year: 1996/97
- Sir Matt Busby Player of the Year: 1996/97
- 1998 FIFA World Cup Team of the Tournament
- UEFA Club Player of the Year: 1999
- Premier League 10 Seasons Awards (1992/3 - 2001/2)
  - Overall Team of the Decade
  - Goal of the Decade (17 August 1996)
- BBC Sports Personality of the Year: 2001

Figure 2. Temporal facts in infobox and honours list

For illustration, Figure 2 lists the infobox and the honours list of David Beckham. For example, "Senior career" attributes are extracted as *playsForSeniorClub* facts, yielding the facts "David Beckham *playsForSeniorClub* Manchester United" and "David Beckham *playsForSeniorClub* Real Madrid". YAGO can accept only one of these two fact candidates, as it enforces consistency constraints like a functional dependency from SoccerPlayer to SoccerClub (a player cannot simultaneously play for two clubs). The reason for this deficiency is that YAGO lacks temporal information. In *T-YAGO*, we can accept both facts if we can validate that they refer to disjoint time periods. We identify time points or intervals like 1993-2003 by pattern matching. This yields temporal facts like "#3: DavidBeckham *playsForSeniorClub* ManchesterUnited", "#3 *since* 1993", and "#3 *until* 2003" (with simplified notation, leaving out the [earliest, latest] part for brevity). Other time-annotated attributes are extracted in a similar way. From the honours lists we also obtain valuable knowledge such as "#4: DavidBeckham *hasWon* BBC Sports Personality of Year" and "#4 *on* 2001".



Categories: 1998 FIFA World Cup players | 2002 FIFA World Cup players | 2006 FIFA World Cup players | 20th-century English people | 21st-century English people | A.C. Milan players | BBC Sports Personality of the Year winners | British expatriate sportspeople in the United States | … | People from Leytonstone | Premier League players | Preston North End F.C. players | Real Madrid C.F. players | Serie A footballers | The Football League players | UEFA Euro 2000 players | UEFA Euro 2004 players | 1975 births | Living people

Figure 3. Temporal facts from categories of David Beckham.

The category systems for Wikipedia articles also contain temporal facts like those highlighted in Figure 3. We identify time points in category names and map them onto the timeline. The remaining part of the category name is treated as the entity. For example, from the first category in Figure 3 we extract the temporal fact "#5: David Beckham *playsAs* FIFA World Cup players", "#5 *on* 1998".

2.3 Query Processing

T-YAGO provides a time-aware query language on its knowledge base of temporal facts. Recall that facts are represented as subject-property-object triples, SPO triples for short, of the RDF data model. Conditions on S, P, or O are expressed by SPARQL triple patterns, and can be combined in a conjunctive manner. For temporal conditions we have extended the query language by a suite of time predicates that refer to the *on*, *since*, and *until* relations of temporal facts:

before, after, equal, during, overlaps, sameYear,

and a few more. Each of these predicates takes as input two time points or two time periods or a time point and a time period, and returns a Boolean value. As *T-YAGO*'s notion of validity times refers to fact identifiers, we need to be able to have variables for fact identifiers and also variables that denote time points for which we need to compute appropriate bindings.

As an example, consider a query about teammates of David Beckham – soccer players who played for the same club as Beckham during overlapping periods. We can express this in the *T-YAGO* quey language as follows:

?id1: "David Beckham" *playsForClub* ?x .
?id2: ?a *playsForClub* ?x .
?id1 *since* ?t1 . ?id1 *until* ?t2 .
?id2 *since* ?t3 . ?id2 *until* ?t4 .
[?t1-?t2] *overlaps* [?t3-?t4] .
?a *notEqual* "David Beckham"

where [?t1-?t2] denotes a time interval. The query returns important players such as Paul Scholes, Gary Neville, and Ruud van Nistelrooy at Manchester United, and Zinedine Zidane, Luis Figo, and Ronaldo at Real Madrid.

In the query, predicates like *overlaps* are used as if they were relations (or properties in RDF jargon). However, they are not necessarily materialized, but instead computed dynamically as needed. We call these virtual relations. As they are actually evaluated as run-time functions, it is easy to switch to *relaxed-matching semantics* as an alternative to exact-matching evaluation of time predicates. The rationale for this option is that we may have different time resolution or uncertainty in the validity times of different facts. For example, consider a query about politicians who visited the same city on the same day. We may know that one politician visited Rome on May 21, and that another politician visited Rome in May of the same year. These time points do not match exactly because of the different resolutions. But we should still consider them as equal in a relaxed-matching mode.

For such cases, *T-YAGO* supports relaxed-matching variants of all temporal predicates. These are based on the [earliest, latest] representation of uncertain time points. Two time points with uncertainty are considered equal if there is a non-zero probability that they are truly equal if they were exactly known. Figure 4 illustrates the matching of time points with [earliest, latest] uncertainty in a graphical manner: while the time points t1 and t2 are not equal in the strict sense they should be regarded as potentially equal applying our relaxed matching scheme**.** The other temporal predicates like *overlaps*, *during*, etc., are handled analogously.

t1 [1999-01-01, 1999-06-30]   t2 [1999-03-01, 1999-12-31]

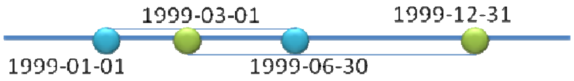
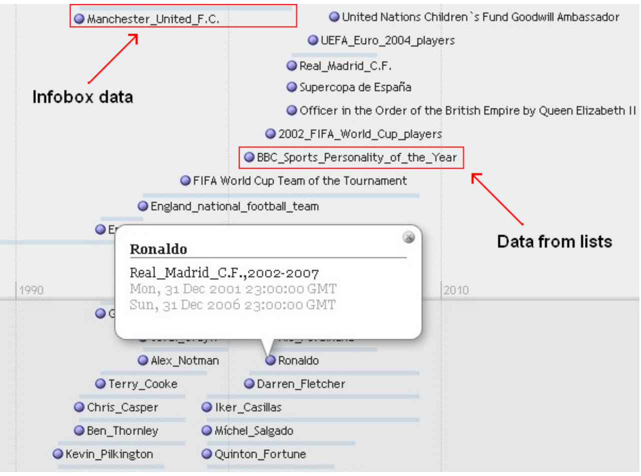


Figure 4. Relaxed matching of time points.

3. DEMO

At this point, *T-YAGO* contains around 300.000 temporal facts from the sports domain. Among them, about 70.000 facts have been extracted from Wikipedia categories and lists embedded in articles. All these temporal facts have been integrated into the existing YAGO knowledge base. This way, we can demonstrate our querying capabilities for temporal facts in our prototype system.


Manchester_United_F.C.
United Nations Children`s Fund Goodwill Ambassador
UEFA_Euro_2004_players
Infobox data
Real_Madrid_C.F.
Supercopa de España
Officer in the Order of the British Empire by Queen Elizabeth II
2002_FIFA_World_Cup_players
BBC_Sports_Personality_of_the_Year
FIFA World Cup Team of the Tournament
England_national_football_team
Data from lists
1990
2010
Alex_Notman
Ronaldo
Terry_Cooke
Darren_Fletcher
Chris_Casper
Iker_Casillas
Ben_Thornley
Míchel_Salgado
Kevin_Pilkington
Quinton_Fortune

Ronaldo

Real_Madrid_C.F.,2002-2007
Mon, 31 Dec 2001 23:00:00 GMT
Sun, 31 Dec 2006 23:00:00 GMT

Figure 5. The timeline of David Beckham.

To present the temporal facts, we implemented our demonstration based on SIMILE Timeline [14] which is a DHTML-based AJAX widget for visualizing time-based events. In our demo, we support several types of temporal queries. For example, users can query temporal facts for a person. Figure 5 is a snapshot segment of the querying result for "David Beckham". A point means a fact valid only on a time point like "David Beckham *wasBornOn* 1975". A time span denotes a fact valid in a time interval, e.g. "David Beckham *playsForSeniorClub* from 1993 to 2003" The users can navigate all facts along the whole timeline and click on anyone of them for the details. Above the timeline are the temporal facts concerning "David Beckham" himself, such as the awards he gained and his duration of time playing for different clubs. Below the timeline, there are the results for the query in subsection 2.3 "who is the teammate of David Beckham?" We can click on each player's name and check his career information when David Beckham was at the same team with him. With the visualization, we can easily identify the related facts if they are in the same or the overlapping time spans.

4. RELATED RESEARCH

There are many existing efforts on automatic ontology construction. Some of them are based on Wikipedia like YAGO[10], Kylin[11], DBpedia[4], and others. They contain millions of entities and tens of millions of relations between entities. Other projects like KnowItAll[6] or TextRunner[5] aim to perform domain-independent and scalable extraction from bigger and noisier web corpora.

The closest work on temporal fact extraction is the event extraction task introduced in SemEval workshops and TimeML [9]. TARSQI toolkit [12] is the state-of-the-art tool for this natural-language processing task. Here, events are defined as tensed verbs, adjectives, and nominals that describe the temporal aspects of the events. For example, adverbial phrases such as "a week ago" or "last year" should be annotated by TARSQI. This task is very different from our crisp notion of temporal facts referring to entities and relations. The recognition and normalization of temporal expressions in natural language are mainly based on linguistic parsing and machine learning [2, 7] or a combination of machine learning and rule-based approaches [1, 8]. TimeBank [13] and TERN 04 [15] annotated corpora are popular training and test sets. There is also recent awareness of Temporal IR [3].

5. OUTLOOK

In this paper we have introduced *T-YAGO*, a fully implemented prototype system which enhances facts extracted from Wikipedia with temporal validity information. While our initial harvesting of temporal facts has focused on the sports domain, we have started to extract facts about politicians, business people, and companies as well. This way we are building a large-scale temporal information system that will be able to trace knowledge evolution over time.

6. REFERENCES

[1] D. Ahn, S. F. Adafre and M. d. Rijke. Towards Task-Based Temporal Extraction and Recognition. In Proceedings of the Dagstuhl Seminar. 2005.

[2] D. Ahn, J. V. Rantwijk and M. d. Rijke. A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. In Proceedings NAACL-HLT 2007.

[3] O. Alonso, M. Gertz, and R. Baeza-Yates. On the Value of Temporal Information in Information Retrieval. SIGIR Forum, December 2007.

[4] S. Auer, C. Bizer, G.Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives. DBpedia: A nucleus for a Web of open data., In Proceedings of the Sixth International Semantic Web Conference (Pusan, Korea, Nov. 11--15). Springer, Berlin/Heidelberg, 2007.

[5] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, O. Etzioni: Open Information Extraction from the Web. IJCAI 2007.

[6] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. 2004. Web-scale information extraction in knowitall: (preliminary results). In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM, New York, NY.

[7] O. Koomiyets, M. Moeans. Meeting TempEval-2: Shallow Approach for Temporal Tagger. In Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 52-57, Boulder, Colorado, June 2009.

[8] C. Min, M. Srikanth and A. Fowler. LCC-TE: A Hybrid Approach to Temporal Relation Identification in News Text. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, June 2007.

[9] J. Pustejovsky, J. Castano, R. Ingria and R. Sauri. TimeML: Robust Specification of Event and Temporal Expressions in Text. In Proceedings of the 5th International Workshop on Computational Semantics. 2003.

[10] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge - unifying WordNet and Wikipedia. Proceedings of WWW07, 2007.

[11] F. Wu and D. Weld. Autonomously semantifying Wikipedia. In Proceedings of the 16th ACM Conference on Information and Knowledge Management (Lisbon, Nov. 6–10). ACM Press, NewYork, 2007.

[12] Q. Zhang, F. M. Suchanek, L. Yue, and G. Weikum. TOB: Timely ontologies for business relations. In Proceedings of the International Workshop on the Web and Databases (WebDB), 2008.

[13] http://www.timeml.org

[14] http://www.simile-widgets.org/timeline

[15] http://fofoca.mitre.org/tern.htm