



max planck institut
informatik

Tiwiki: Searching Wikipedia with Temporal Constraints

Prabal Agarwal and Jannik Strötgen

`jannik.stroetgen@mpii.de`

TempWeb'17 – Perth, Australia – April 3, 2017

Temporal Information Retrieval (TIR) – a well established field

■ challenges and opportunities

- Alonso et al.:
Temporal Information Retrieval: Challenges and Opportunities.
In TempWeb, 2011.

■ survey

- Campos et al.:
Survey of Temporal Information Retrieval and Related Applications.
ACM Computing Surveys, 47(2):15:1–15:41, 2014.

■ general overview

- Kanhabua et al.:
Temporal Information Retrieval.
Foundations and Trends in Information Retrieval, 9(2):91–208, 2015.

TIR – several directions

■ **time as dimension of relevance**

- exploited: document creation times
- by identifying time interval of interest, documents from that time can be selected
- freshness of documents for news-related queries

■ **time as context information**

- exploited: time of query
- depending on query time, same queries represent different information needs

■ **time as constraint about the content**

- exploited: temporal information in the documents' content
- what time interval should results be talking about

TIR – time as constraint about the content

types of queries

- queries with non-explicit temporal constraints:
 - atemporal, temporally ambiguous, temporally unambiguous [Jones & Diaz, 2007]
- queries with explicit temporal constraints

importance

- 13.7 % of queries are explicit temporal queries [Zhang et al., 2010]

most similar to our work

- approaches with explicit temporal constraints: $q = \langle q_{text}, q_{time} \rangle$
- several approaches – in research
[Berberich et al., 2010; Strötgen/Gertz, 2013, ...]

TIR – time as constraint about the content

today's Web search engines

- no (easy) way to formulate queries of the form $q = \langle q_{text}, q_{time} \rangle$
- temporal expressions (in queries and documents) are not well handled

research prototypes

- no dynamic document collections
- no publicly available system for searching with temporal constraints

what's missing

- a publicly available time-aware search engine for a dynamic set of documents

Goal of this Work

to provide a time-aware search engine for a dynamic document collection



<input type="text" value="Search query"/>	<input type="text" value="Start Date"/>	<input type="text" value="End Date"/>
---	---	---------------------------------------

Pre-requirement for “time as constraint about the content”

temporal expressions have to be taken care of

- in queries
- in documents

Temporal Tagging – “to take care of temporal expressions”

1. **extraction** of temporal expressions


Greece Makes ‘Good Progress’ in Payment Talks

By Maria Petrakis and Natalie

Sep 20, 2011 10:39 PM GMT+0200 [ADD TO QUEUE](#)

[Recommend](#) 10 [Tweet](#) 33 [Share](#) 4 [+1](#) 2

[Enlarge image](#)



Evangelos Venizelos, Greece's finance minister, listens to an aide during a meeting.

Tsionis/Bloomberg

Greek Finance Minister Evangelos Venizelos made “good progress” in a second round of talks with the European Union and [International Monetary Fund](#) aimed at staving off default, the EU said.

The telephone meeting **tonight**, which followed **yesterday**, were intended to damp concerns that Greece's debt reduction targets and to clear the way for a sixth installment of rescue funds. The EU statement said a “full mission” will return to Athens **next week** discussions in coming days at the IMF's annual meeting in [Washington](#).

Staying in the euro area is an “irreversible and fundamental national choice,” Venizelos said in a statement. **Earlier today**, he pledged that “our financial data and economic structures are a problem for the euro zone, but we are determined to tackle once and for all.”

The EU's next payment for Greece is on track to be released **next month**, minister [George Papandreu](#) fights

Temporal Tagging – “to take care of temporal expressions”

1. **extraction** of temporal expressions
2. **normalization** of temporal expressions

Greece Makes ‘Good Progress’ in Payment Talks

By Maria Petrakis and Natalie

Sep 20, 2011 10:39 PM GMT+0200

ADD TO QUEUE

Recommend 10

Tweet 33

Share 4

+1 2

Enlarge image



Evangelos Venizelos, Greece's finance minister, listens to an aide during a

Tsionis/Bloomberg

Greek Finance Minister Evangelos Venizelos made “good progress” in a second round of talks with the European Union and [International Monetary Fund](#) aimed at staving off default, the EU said.

The telephone meetings **tonight**, which followed **yesterday**, were intended to damp concerns that Greece's reduction targets and to clear the way for a sixth installment of rescue funds. The EU statement said a “full mission” will return to Athens **next week** discussions in coming days at the IMF's annual meeting in Washington.

Staying in the euro area is an “irreversible and fundamental national choice,” Venizelos said in a statement. He **earlier today** pledged that “our financial data and economic structures are a problem for the euro zone, determined to tackle once and for all.”

The EU's next payment for Greece is on track to be released **next month**, minister George Papandreu fights

tonight → **2011-09-20TNI**

yesterday → **2011-09-19**

next week → **2011-W39**

Sept. 20, 2011 → **2011-09-20**

next month → **2011-10**

Pre-requirement – temporal tagging

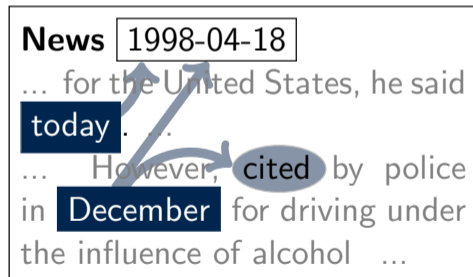
different document types pose different challenges

- normalization of relative / underspecified expressions
- type of language (formal, scientific, historic, colloquial)

News 1998-04-18

... for the United States, he said
today ...

... However, **cited** by police
in **December** for driving under
the influence of alcohol ...



Narrative 2009-12-19

1979 : Soviet invasion
... land in Kabul on
December 25 ...
Afghanistan from the north on
December 27. In **the morning**



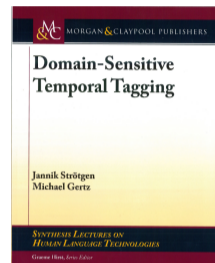
Pre-requirement – temporal tagging

different document types pose different challenges

- normalization of relative / underspecified expressions
- type of language (formal, scientific, historic, colloquial)

domain-sensitivity of a temporal tagger is important

- was ignored for a long time
 - now: first taggers not tailored for news documents only
- book on “domain-sensitive temporal tagging”



news

- news articles
- formal emails

narrative

- Wikipedia
- descriptive docs

colloquial

- SMS, tweets
- user-g. content

autonomic

- scientific doc.
- literary works

The *tiwiki* Search Engine – back to this paper



<input type="text" value="Search query"/>	<input type="text" value="Start Date"/>	<input type="text" value="End Date"/>
---	---	---------------------------------------

The *tiwiki* Search Engine – characteristics

special type of query $q = \langle q_{text}, q_{time} \rangle$

- with $q_{time} = [begin, end]$
- *begin* and *end* with day, month, or year granularity
- ranking based on $score = \alpha \cdot score_{text} + (1 - \alpha) \cdot score_{time}$

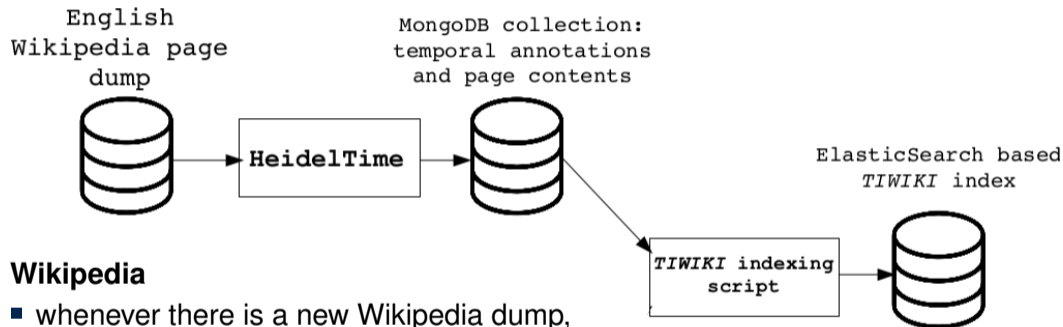
dynamic document collection: Wikipedia

- steadily updated
- frequently accessed
- heuristic observation as motivation:
many Web queries result in at least one Wikipedia page within top ranks

publicly available

- tiwiki.mpi-inf.mpg.de
- tiwiki.wannauchimmer.de

The *tiwiki* Search Engine – indexing pipeline



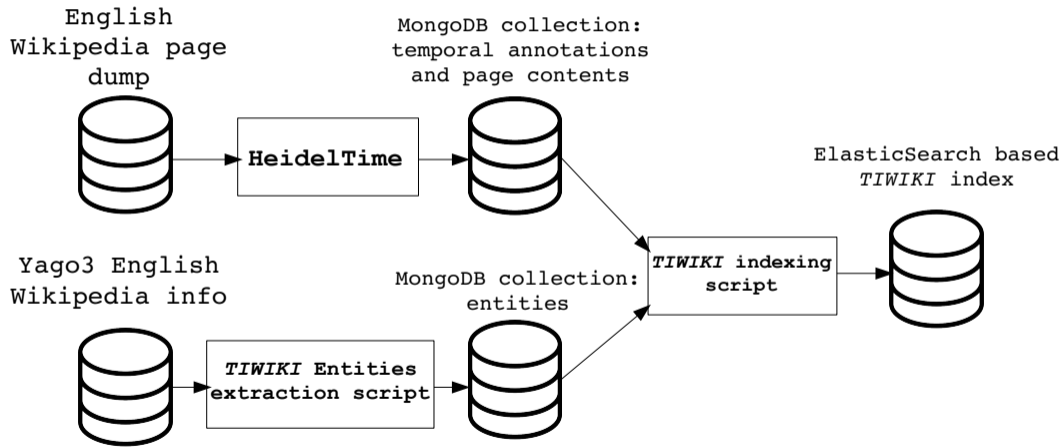
Wikipedia

- whenever there is a new Wikipedia dump, *tiwiki* gets updated

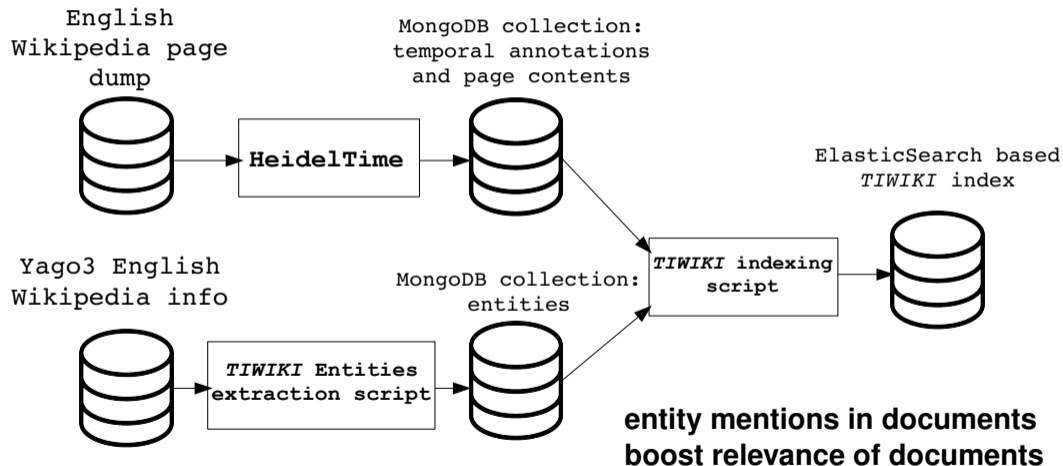
HeidelTime

- dumps are processed with the domain-sensitive temporal tagger HeidelTime (also publicly available)

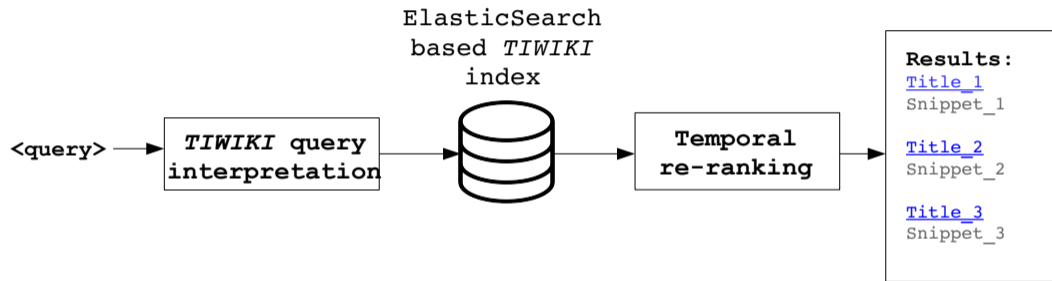
The *tiwiki* Search Engine – indexing pipeline



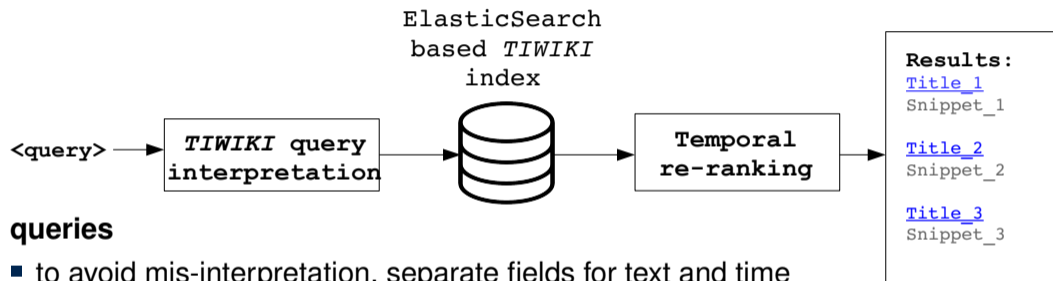
The *tiwiki* Search Engine – indexing pipeline



The *tiwiki* Search Engine – querying pipeline



The *tiwiki* Search Engine – querying pipeline



queries

- to avoid mis-interpretation, separate fields for text and time

temporal snippets

- keyword matches are highlighted
- temporal expressions matching the temporal constraints are highlighted
- normalized values of temporal expressions are shown additionally

Evaluating *tiwiki*

Evaluation of temporal IR

- hardly any well fitting benchmarks for temporal IR
- if queries with temporal constraints, benchmarks tailored for news collections (e.g., Temporalia)

Berberich et al. (2010)

- used MTurk to get a set of meaningful queries with “text and time”
- temporal constraints of different granularities
 - day, month, year, decade, century
(though only single expression “time intervals”)
- different topics
 - sports, technology, world affairs, culture
- developed to evaluate on a news archive, but best suited queries for *tiwiki*

Evaluating *tiwiki*

Goal of the evaluation

qualitative not quantitative demonstration of *tiwiki*'s usefulness

comparison of top ranked Wikipedia articles

- q_{text} = keywords + standard English writing of temporal expression, $q_{time} = []$
- q_{text} = keywords, q_{time} = normalized value of temporal expression

Evaluating *tiwiki* – Some Examples

$q_{text} = \text{'ac milan may 23 2007'}$, $q_{time} = []$

versus

$q_{text} = \text{'ac milan'}$, $q_{time} = [2007-05-23, 2007-05-23]$

Activities Firefox ESR Sun 18:03

tiwiki - Mozilla Firefox (Private Browsing)

tiwiki

tiwiki.mpi-inf.mpg.de

Search

tiwiki

time-aware search for Wikipedia

ac milan may 23 2007 Start Date End Date

Search

[Milan](#)

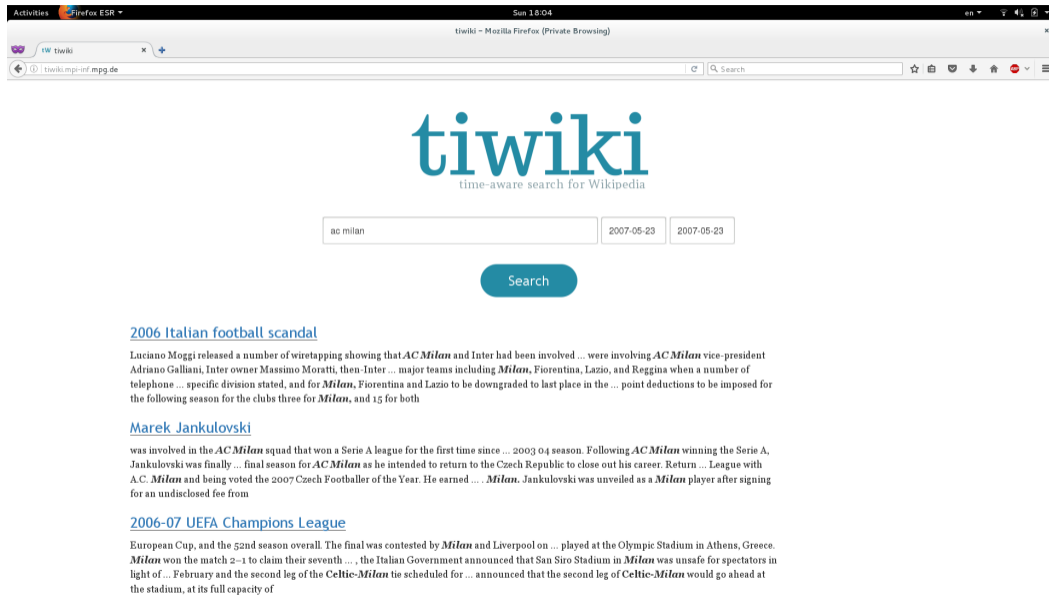
football teams, *AC Milan* and FC Internazionale. Toponymy. The etymology of *Milan* Lombard: "*Milan* ... "Teneke", by Fabio Vacchi in *2007*. Other major theatres in *Milan* include the Teatro ... *Milan* or ; ; Lombard, Milanese variant: "*Milan*", is the capital of the Lombardy region, and the The wider *Milan* metropolitan area, known as "Greater *Milan*", is a polycentric metropolitan region of 7 ... to 10 million people. *Milan* is the main industrial and financial centre of Italy and one of global

[War of Currents](#)

later by an *AC* line, and a Brush Electric Company lineman killed in *May* ... and US Supreme Court denied Kemmler's appeal on *May 23*, 1890 ... Light Company and the alternating current *AC* based Westinghouse Electric Company. It took place ... by accidental electrocution from high voltage *AC* leading to a debate over its safety and regulation ... that directly competed for the market the Edison system was designed to serve. As many other *AC*

[Lockheed AC-130](#)

44 Bushmaster II cannons. In *2007*, the Air Force modified four *AC-130U* gunships as ... armament *may* only be installed on a few aircraft rather than the entire *AC-130J* fleet; the laser ... 15 *May* 1975 : Six *AC-130s* and 52 air crew members were lost during the war ... , infrared and radar sensors. In *2007*, US Special Operations forces also used the *AC* ... The Lockheed *AC-130* gunship is a heavily armed, long-endurance ground-attack variant of the C-130



Activities Firefox ESR Sun 18:04

tiwiki - Mozilla Firefox (Private Browsing)

tiwiki

tiwiki.mpi-inf.mpg.de

Search

tiwiki

time-aware search for Wikipedia

ac milan 2007-05-23 2007-05-23

Search

[2006 Italian football scandal](#)

Luciano Moggi released a number of wiretapping showing that *AC Milan* and Inter had been involved ... were involving *AC Milan* vice-president Adriano Galliani, Inter owner Massimo Moratti, then-Inter ... major teams including *Milan*, Fiorentina, Lazio, and Reggina when a number of telephone ... specific division stated, and for *Milan*, Fiorentina and Lazio to be downgraded to last place in the ... point deductions to be imposed for the following season for the clubs three for *Milan*, and 15 for both

[Marek Jankulovski](#)

was involved in the *AC Milan* squad that won a Serie A league for the first time since ... 2003/04 season. Following *AC Milan* winning the Serie A, Jankulovski was finally ... final season for *AC Milan* as he intended to return to the Czech Republic to close out his career. Return ... League with A.C. *Milan* and being voted the 2007 Czech Footballer of the Year. He earned ... *Milan*. Jankulovski was unveiled as a *Milan* player after signing for an undisclosed fee from

[2006-07 UEFA Champions League](#)

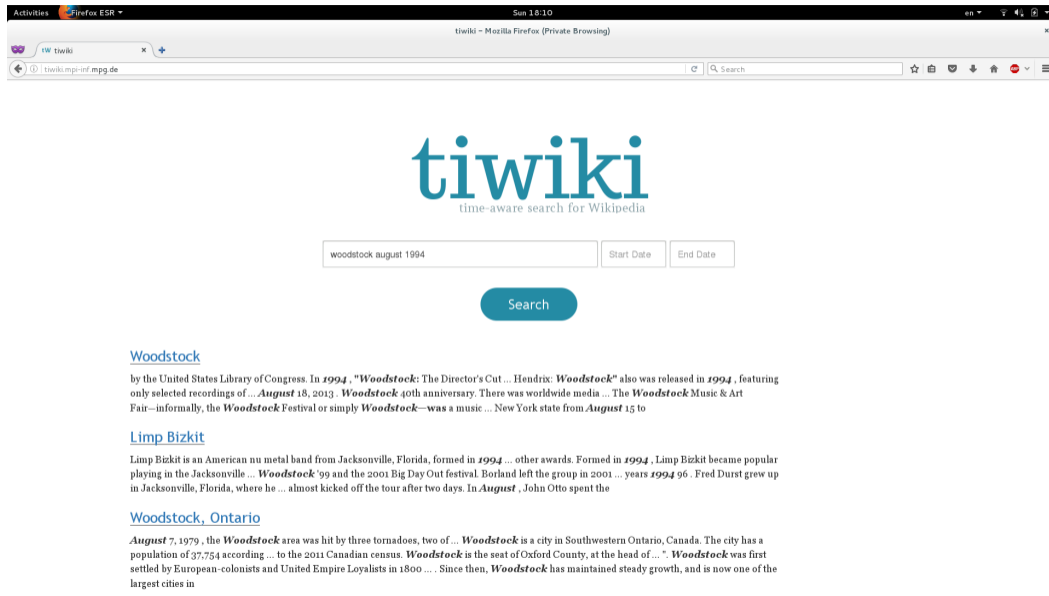
European Cup, and the 52nd season overall. The final was contested by *Milan* and Liverpool on ... played at the Olympic Stadium in Athens, Greece. *Milan* won the match 2–1 to claim their seventh ... , the Italian Government announced that San Siro Stadium in *Milan* was unsafe for spectators in light of ... February and the second leg of the Celtic-*Milan* tie scheduled for ... announced that the second leg of Celtic-*Milan* would go ahead at the stadium, at its full capacity of

Evaluating *tiwiki* – Some Examples

$q_{text} = \text{'woodstock august 1994'}$, $q_{time} = []$

versus

$q_{text} = \text{'woodstock'}$, $q_{time} = [1994-08, 1994-08]$



The screenshot shows a web browser window with the title "tiwiki - Mozilla Firefox (Private Browsing)". The address bar shows the URL "twiki.mpi-inf.mpg.de". The page features the "tiwiki" logo with the tagline "time-aware search for Wikipedia". Below the logo is a search interface with a text input field containing "woodstock august 1994", and buttons for "Start Date", "End Date", and a "Search" button. The search results are displayed below the input field, showing three entries: "Woodstock", "Limp Bizkit", and "Woodstock, Ontario".

Woodstock

by the United States Library of Congress. In **1994**, "**Woodstock**: The Director's Cut ... Hendrix: **Woodstock**" also was released in **1994**, featuring only selected recordings of ... **August** 18, 2013 . **Woodstock** 40th anniversary. There was worldwide media ... The **Woodstock** Music & Art Fair—informally, the **Woodstock** Festival or simply **Woodstock**—was a music ... New York state from **August** 15 to

Limp Bizkit

Limp Bizkit is an American nu metal band from Jacksonville, Florida, formed in **1994** ... other awards. Formed in **1994**, Limp Bizkit became popular playing in the Jacksonville ... **Woodstock** '99 and the 2001 Big Day Out festival. Borland left the group in 2001 ... years **1994** 96 . Fred Durst grew up in Jacksonville, Florida, where he ... almost kicked off the tour after two days. In **August**, John Otto spent the

Woodstock, Ontario

August 7, 1979, the **Woodstock** area was hit by three tornadoes, two of ... **Woodstock** is a city in Southwestern Ontario, Canada. The city has a population of 37,754 according ... to the 2011 Canadian census. **Woodstock** is the seat of Oxford County, at the head of ... ". **Woodstock** was first settled by European-colonists and United Empire Loyalists in 1800 Since then, **Woodstock** has maintained steady growth, and is now one of the largest cities in

Activities Firefox ESR Sun 18:11

tiwiki - Mozilla Firefox (Private Browsing)

tiwiki

tiwiki.mpi-inf.mpg.de

Search

tiwiki

time-aware search for Wikipedia

woodstock 1994-08 1994-08

Search

[Woodstock '94](#)

Woodstock '94 was a music festival organized to commemorate the 25th anniversary of the original ... **Woodstock** festival of 1969 . It was promoted as "2 More Days of Peace and Music." The ... north of New York City. The site is from **Woodstock**, New York. The crowd at **Woodstock** '94 was ... safety, security and peaceful atmosphere. The festival was followed by the infamous **Woodstock** 1999 ... **Sunday, August 13** [1994-08] and **14** [1994-08]

[Woodstock Jimi Hendrix album](#)

Woodstock is a live album by Jimi Hendrix released posthumously on [1994-08] ... Gypsy Sun and Rainbows at **Woodstock** Festival, Bethel, New York on ... as "Live at **Woodstock**", albeit with some of the performers mixed out. Critical reception. In a ... , highly eccentric music on "**Woodstock**" makes for what was a transitional but all-important live album by ... happen any more? Not on such a scale for sure." Although "Live at **Woodstock**" 1999 was

[Michael Wadleigh](#)

groundbreaking documentary of the 1969 **Woodstock** Festival, "**Woodstock**". Biography. A native of ... , Wadleigh undertook the monumental task of documenting the **Woodstock** Music Festival. He arrived on ... profits. "**Woodstock**" stands as a milestone in the documentary film field, receiving an Academy Award for ... be eleven years after the release of "**Woodstock**" before he received his next and, last to date ... showing, a 228-minutes "director's cut" of "**Woodstock**" was released, and in 1999

Evaluation – *tiwiki* for named entity disambiguation

$q_{text} = \text{"george bush"} , q_{time} = []$

1. George W. Bush
2. George H. W. Bush

$q_{text} = \text{"george bush"} , q_{time} = [1990, 1993]$

1. George H. W. Bush
2. Presidency of George H. W. Bush

$q_{text} = \text{"george bush"} , q_{time} = [2001, 2009]$

1. George W. Bush
2. Public image of George W. Bush

Summary & Ongoing Work

tiwiki

- publicly available search engine for Wikipedia
- handles queries with temporal constraints

things to do

- constantly maintain *tiwiki* to guarantee up-to-date results
- multilingual extension of *tiwiki* (HeidelTime is multilingual)
- improve the ranking and add features
(e.g., textual proximity of matched keywords and temporal expressions)

Thank you – and please satisfy your temporal information needs with *tiwiki*!



<input type="text" value="Search query"/>	<input type="text" value="Start Date"/>	<input type="text" value="End Date"/>
<input type="button" value="Search"/>		

tiwiki.mpi-inf.mpg.de
or
tiwiki.wannauchimmer.de