SAWUS
Siena's Automatic Wikipedia Update System

Carl Tompkins, Zachary Witter and Sharon G. Small

The Siena College Institute for Artificial Intelligence
Siena College
Loudonville, NY 12211
{cl05tomp, zl14witt, ssmall}@siena.edu

Abstract:
The National Institute of Standards and Technology (NIST) has been running an annual Text Retrieval Competition and Conference (TREC) since 1992. This is a premier conference that offers researchers in the field of Computational Linguistics the opportunity to showcase their work and compare their results against other leading researchers. Our Siena research team participated in the TREC Knowledge Based Acquisition (KBA) Track, which was offered for the first time in 2012. The objective of this track is to drive research into automatic acquisition of knowledge, such as automatically updating Wikipedia by utilizing online news. Specifically our team of researchers developed a system that filters a stream of content for information that should be included on a given Wikipedia page. It was not yet clear how traditional Information Retrieval (IR) techniques perform for this task; therefore we began with a baseline test using current state of the art IR techniques. We then went on to experiment with query expansion, building a module that utilized Wikipedia Infoboxes to add terms to our query. This module was incorporated with our IR component to create SAWUS. Four submissions were sent to NIST to undergo a formal evaluation.

Introduction

NIST selected Wikipedia (WP) as their source for a knowledge base (KB). Participating teams were provided with a content stream, ~500,000 English articles from online news, blogs and forums. This set of data was used to simulate real world conditions, that is, a steady content stream of data. Automatically processing this content stream, systems were required to identify relevant content and then recommend edits to the proper corresponding Wikipedia pages. Where the final step being that human curators would be able to utilize these suggestions to make faster and more accurate updates to a knowledge base, Wikipedia. The ultimate objective of the SAWUS project was to provide a completely real time system that takes a steady input stream of internet data, indexes it, and compares a ranked list of documents to an existing Wikipedia page to judge the material's relevancy.

This problem posed significant challenges to researchers:

1. While many news articles explicitly mention entities with WP nodes, some relevant articles do not. Detecting these articles and linking them to appropriate WP nodes will require more sophisticated filtering techniques than simple name matching.

2. Important attributes of an entity might change before a human editor assimilates this new information into the KB. Thus, a KBA system may need to refer to its own proposed-but-not-yet-accepted edits in order to properly assess subsequent items in the stream.
3. Only after sufficient evidence has accumulated should the system summon the scarce resources of human curators. How can one define "sufficient" evidence?

Preparing the Data Collection

The SAWUS corpus consisted of news, social media, and bit.ly links taken from numerous websites. The data was scraped from these types of websites by NIST. This process was done from October 7, 2011 through May 1, 2012. Three different sets of corpora were available to download: a 1.9TB, a 1.1TB, and a 275GB version. Each version of the corpus contained lessening amounts of data, but still covering all of the dates. The 1.9TB corpus is the full corpus including raw HTML text. The 1.1TB corpus includes only documents marked as possibly being in English and the raw HTML text removed, but still contains named entity tagged data. The 275GB also contains English-only documents and cleansed text, but does not contain any named entity tagged data. The data was available for download in numerous ways, including Amazon S3, which is an online service offered by Amazon that allows users to store large amounts of data. We downloaded the corpus via HTTP and chose to use the 275GB corpus, mainly because of size limitations. This corpus consisted of 4,973 directories, each named for the date and the hour of when the content was retrieved. There were 352,861 files distributed through the directories, each with a .xz.gpg extension. Each file was encrypted due to the public availability of the corpus download and the requirement to agree to the TREC non-disclosure agreement. In order to use the data, we needed to decrypt all of the files using the provided key; we accomplished this task by writing a bash script to iterate over each directory. Since the total corpus was 275GB and our space was limited, we decided to keep the files compressed and decompress the files at indexing time. We examined part of the social data and determined that it was not very useful for our task. The social data was difficult for the HTML cleanser to process, resulting in an overzealous removal of text from the data. With very little and unhelpful text, we removed the social data from the corpus to reduce the amount of data to index. The final size was 185GB of data.

SAWUS

The current system's core functionality is built around Apache's powerful indexing technology known as the *Apache Lucene Core* or simply *Lucene*. SAWUS accepts the KBA topics in the TREC predefined format. Our Query Processor module extracts the entity name, i.e. *Ahron Barak, Bill Coen, etc*. and the time stamp. The entity name is required to find the proper Wikipedia page but also it is used by Lucene in order to retrieve a relevant set of articles from the news stream. For our baseline, we just returned this ordered list of relevant articles without further processing. In further experimentation, we created a Query Expansion Module in an attempt to improve the ranking of our results.

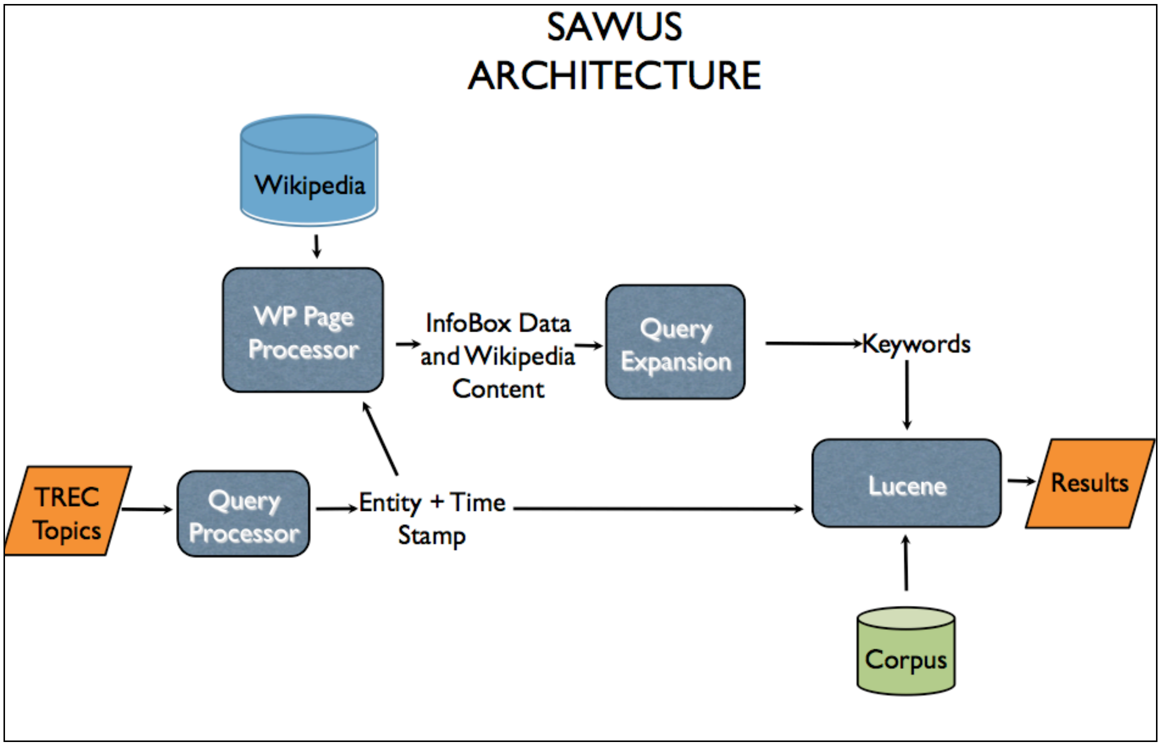
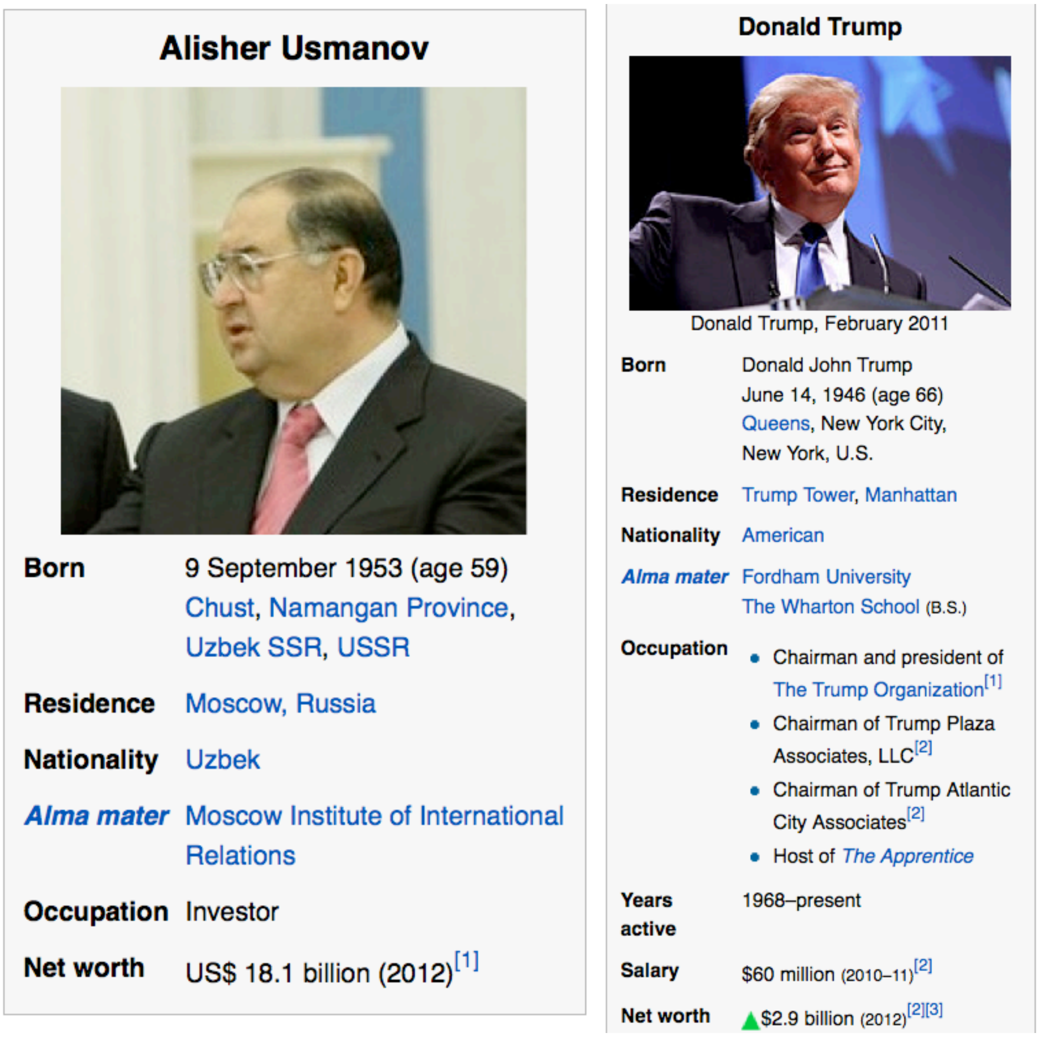


Figure 1: SAWUS Architecture Diagram

Query Expansion

Our first step in Query Expansion (QE) was to utilize the entity name from the KBA topics to automatically access Wikipedia and retrieve the corresponding page for the input entity. This step was simple as the entity name was the exact suffix of the url for the entity's Wikipedia page. We then used the input timestamp to automatically search through the wikipedia history page of the specified entity and find the exact wikipedia page that would have existed at the given time. Our QE module then extracted the text from the page and parsed the information box (infobox). Nearly every Wikipedia page contains an infobox, Figure 2. Note the semi-structured format of these boxes. We hypothesized that extracting information from these boxes and utilizing the text to expand on the original query will improve the results we retrieve from the content stream. For example utilizing information from the infoBox of "*George Washington*" expands the query from "*George Washington*" to ("*George_Washington" AND ("John Adams" OR President OR "United States*"...). The "AND" requirement requires one of these expanded terms to exist in the documents retrieved by Lucene, which eliminates some of the ambiguous pages we were retrieving, e.g. "*George Washington University*".



Alisher Usmanov

Born	9 September 1953 (age 59) Chust, Namangan Province, Uzbek SSR, USSR
Residence	Moscow, Russia
Nationality	Uzbek
Alma mater	Moscow Institute of International Relations
Occupation	Investor
Net worth	US$ 18.1 billion (2012)[1]

Donald Trump

Donald Trump, February 2011

Born	Donald John Trump June 14, 1946 (age 66) Queens, New York City, New York, U.S.
Residence	Trump Tower, Manhattan
Nationality	American
Alma mater	Fordham University The Wharton School (B.S.)
Occupation	• Chairman and president of The Trump Organization[1] • Chairman of Trump Plaza Associates, LLC[2] • Chairman of Trump Atlantic City Associates[2] • Host of *The Apprentice*
Years active	1968–present
Salary	$60 million (2010–11)[2]
Net worth	▲$2.9 billion (2012)[2][3]

Figure 2: Example Information Boxes from Wikipedia

Results

Unfortunately late system hardware issues hindered us in submitting the proper runs for evaluation.