CAKES: Crowdsourced Automatic Keyword Extraction

Daniel Erenrich
erenrich@caltech.edu

Chris Kennelly
ckennelly@ugcs.caltech.edu

ABSTRACT

We present progress towards applying "Games With A Purpose" (GWAP) techniques to extract keywords which describe films. We discuss the use of machine learning algorithms to automatically extract keywords from movie scripts. The data collected from the game is used to generate similarity metrics between the films which eventually can be used to recommend films. Finally, we discuss the merits of GWAP as a system and make recommendations concerning its use.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis; H.5.2 [**User Interfaces**]: Natural language

General Terms

Measurement

1. MOTIVATION

As part of the CS144 "Rankmaniac 2010" competition, we created an online game using the Internet Movie Database (IMDB) dataset. Hoping that it would encourage incoming links, we chose to make a game so our website would have content. The initial game gave the player a set of tags for a randomly chosen movie in the dataset. The goal of the player was to guess the movie possessing those tags in sixty seconds. If needed, the player could request more tags from the pool. From the start of the project, we chose to diligently log game play in order to construct a fascinating dataset for movies.

This strategy has been dubbed by Luis von Ahn, a professor at Carnegie Mellon University, a "Game with a Purpose." These games serve to be both entertaining while nevertheless guiding players to simultaneously solving difficult problems for machines. For example, he designed an image labeling game which asks two players to describe an image with the same word.

Sixty-five years ago, Vannevar Bush presented a vision for an automated system of data organization entitled "the Memex" [1]. Since then, the task of automatically organizing, summarizing, and indexing information still remains. While search engines perform much of the task of indexing, automatic summarization is a work in progress [2]. Like image recognition, keyword assignment to text challenges machines, both in describing a large piece of content with relevant keywords as well as ascertaining the importance of each keyword. This task is important for content recommendation systems and search relevance.

Game-playing humans provide a potential audience for this task. The dataset produced by gameplay can be used to assess the relatedness of films and quality of keywords describing films. Additionally, possible keyword generation algorithms can be validated on-the-fly by emulating an opponent. Rather than summarize widely available texts, we chose keyword generation. In order to apply the Games With A Purpose-style approach to validation, we must keep our game interesting in order to encourage play. Summarizing dusty tomes into multiple paragraphs may be an interesting AI problem but lacks the pace that keywords appear to provide to attract interest as a fun game to a broad audience.

2. METHODOLOGY

Our research focuses on automatic keyword generation from underlying source texts and keyword validation by human game play. The former produces keywords by analyzing source material, movie scripts, automatically. The latter maximizes the value of keywords by attempting to discover which keywords convey the most information to a human about the content of a film by leading them to guess a movie title. Statistical information gleaned from the work in keyword presentation provides feedback to keyword discovery algorithms for selecting more optimal keywords.

Gameplay lends itself to several potentially noisy approaches for measuring the usefulness of keywords for a human player. We considered focusing on the last displayed keyword, assigning equal weight to each displayed keyword, and giving an exponential decay of weights skewed towards more recently displayed keywords. For a "stuck" user, the last displayed keyword ought to give the most insight into the movie of interest, justifying the exclusive use of last keyword for measuring effectiveness. However, many keywords are used to describe multiple movies, limiting the usefulness of this

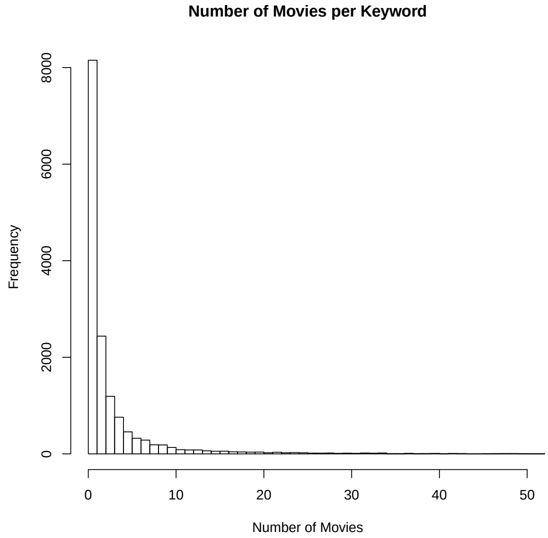


Figure 1: Distribution of frequency of keyword appearances

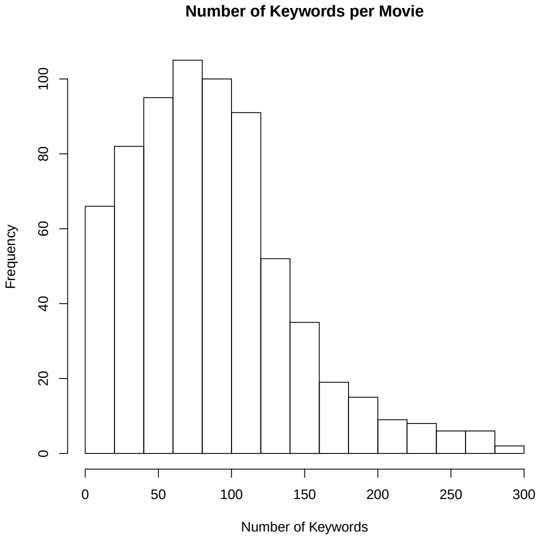


Figure 2: Distribution of frequency of keyword counts

technique. This suggests that the whole set of keywords is greater than the sum of its parts.

The other aspect of the project is the games with a purpose angle. We are using techniques similar to games distributed on the Games With A Purpose website created by Luis von Ahn [3]. These games are designed to be maximally fun while at the same time extracting usable information. There are both one player and two player games on the GWAP website. We implemented a two-player game centered around tagging movies similar to GWAP's "ESP." ESP works by having two users describe a particular image and the round ends when both users choose the same word for the image and a new image is shown. Additionally, some terms are marked as "taboo" and cannot be used. Taboo words are used to ensure that new rounds of the game do not result in keywords that are already known. In our version, one user describes the film and the other user guesses the movie's name. This alteration allows us to insert a "bot" into one half of the game in order to test new keyword extraction algorithms and to provide an opponent when no opponent is available. Our two-player game is a combination of the so-called "output-agreement" and "input-agreement" games which is called an "inversion problem game" [4]. The one player version of the game is the game that already exists and is used to provide humans with game play when no human is available to be matched. Users are shown a series of keywords and they then attempt to use to use them to guess the film's name. We use the frequency with which certain words result in correct identification to determine how likely words are to be good descriptions of a film.

3. DATASETS

3.1 The IMDB Dataset

The Internet Movie Database (IMDB) releases its underlying data for non-commercial use. Figures 1 and 2 give distributions on the frequencies of keywords as associated with movies.

3.2 Movie Scripts

Scripts were scraped from The Internet Movie Script Database (IMSDb). After being downloaded, they automatically cleaned for HTML markup. While some formatting could provide additional information to indicate distinctions between stage directions, scenes, and dialogue, we felt that accurately distinguishing between these across hundreds of scripts would be difficult. Small variations could easily disrupt attempts reliably analyze this information.

Ultimately, 845 scripts were collected for analysis. Manual examination showed that there was considerable variation between some scripts and their respective movies in the form of scene additions and deletions. Figure 3 shows the distribution of unique words in each script. Figure 4 shows the frequency of the number of movie appearances each word had.

4. KEYWORD EXTRACTION

As this project sought to extract keywords from movies, we applied two approaches for automatically analyzing scripts.

4.1 Statistical Techniques

In the literature, the most simple and common technique for extracting key words from bodies of text are based on simple word frequency. These methods count the number of times each words appears and then declare a word to be a keyword if it appears very frequently. Obviously we do not want to consider stop-words like "the" or "hello" in our set of keywords. One technique that deals with this problem has a program scan a corpus of text first and then extracts words which are unusually common from a text. That is, the word appears more often than in the average text. Applying this algorithm to the film scripts in our possession revealed

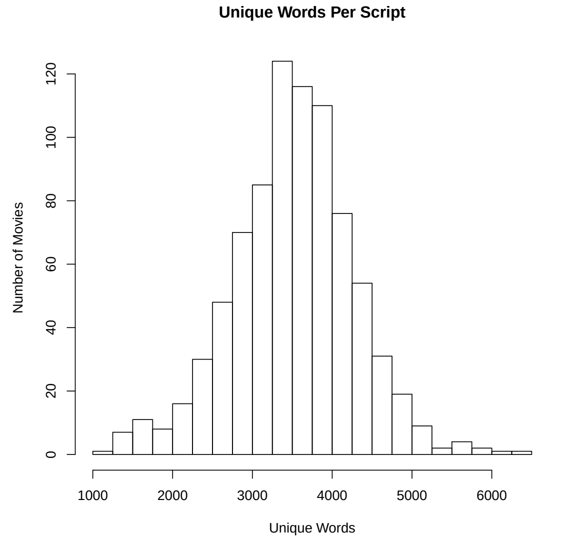


Figure 3: Distribution of Unique Words per Script

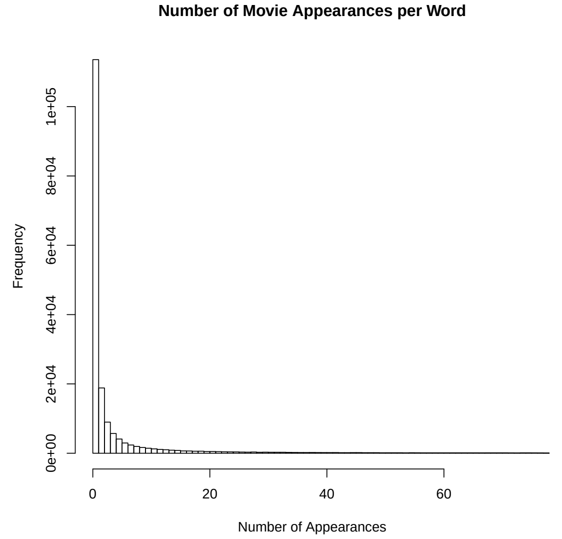


Figure 4: Distribution of frequency of word appearances

reasonable, but not great keywords. An example keyword set for the film "Watchmen" is:

> rorschach, laurie, dan, adrian, blake, manhattan,owl, continued, jon, hollis, moloch, blake's, dr, rorschach's, veidt, moloch's, comedian, janey, dan's, watchmen, ship, karnak, sally,laurie's, forbes, slater, cont'd, det, adrian's, seymour, roth, doug, thug, lynx, dreiberg, nite, industries, manhattan's, jupiter, pyramid, osterman, antarctica, daniel, costume, jon's, mars, psychiatrist, thug's, ion, flashback, agent, frontiersman, swat, gallagher, particles, kashmir, gang, anchorwoman, archie, kovacs, intruder, editor, smartest, mask, prison, cancer, enterprises, edward, rockefeller, vendor, welder, supervillian,

The words are all relevant, but are not what we would consider keywords. Even if we were to remove proper-nouns and stem the remaining words we would be left with a collection of keywords which fail to capture many of the most important characteristics of the film.

From the original IMDB set as well as our game play-based dataset, we assigned keywords to movies. Standard keyword generation algorithms accept blocks of text and find words which have statistically more appearances than might be expected. Film scripts formed the basis for these algorithms as direct analysis of the audio and video of a film is a much more difficult and error-prone task. We considered and tested the use of film-reviews as a corpus, but the results were not promising enough to justify the much larger difficulty in assembling a good corpus.

After manual review, this approach produced generally poor results and were not validated with the crowdsourced system. There is not much hope for these statistical methods to find the sort of higher-order keywords which are often the most useful.

4.2 Machine Learning-Based

Disappointed by the results from applying statistical methods only to scripts, we applied machine learning techniques to map from the features of scripts into a higher-level keyword space. These techniques allow us to escape the literal text of the movie. Even the state-of-the-art in statistical keyword extraction algorithms cannot find keywords such as "Christ analogy" in a source text. Consequently, techniques which leveraged more extensive textual analysis as well as higher-level keywords could be used to escape these limitations.

For this task, we correlated in-script words with the prelabeled IMDB dataset. The dataset collected by the initial run offered some information about the relative usefulness of keywords but was constrained by initial design choices in the game which caused the dataset to be too noisy. The updated game engine saw substantially less traffic during the term and gave an excessively sparse dataset, especially when coupled with user-provided, free-form text.

The size of the underlying feature space derived from scripts greatly constrained the techniques which could be used. As shown in Figure 3, each movie typically had thousands of unique words in its script. Consequently, it was not possible to consider using a tree-augmented naïve Bayes algorithm as the features must be considered *pairwise* and subsequently *stored*. Other algorithms were similarly infeasible due to the size of the dataset being considered.

A breakdown of the predicted keywords for the movie *Enemy of the State* is shown in Table 1. Some keywords may have been erroneously predicted due to differences between the script and the movie's on-screen depiction. The IMSDb script possess a deleted scene in a Catholic church. While such a connection is tenuous, many IMDB keywords are related to films in this way.

For some keywords, the script offered little discernible information which would correlated with the keyword. The keyword "altered version of studio logo" is associated with 62 movies in the IMDB dataset. Using the naïve Bayesian algorithm, 78 movies from the testing set were predicted to be associated with the keyword. Unfortunately, only two of these films matched with the underlying IMDB dataset: *Contact* and *Edward Scissorhands*. Spot checks of the other movies confirmed the accuracy of the IMDB set.

Naïve Bayesian methods faired better with other keywords, highlighting weaknesses within the IMDB labels. Table 2 lists the movies which were labeled with the "surveillance" keyword and whether they were similarly classified within the IMDB set itself. The algorithm correctly labeled *Enemy of the State* and *The Departed*. It classified the James Bond film *Tomorrow Never Dies* with the keyword but failed to do so for its successor *The World Is Not Enough*. Noticeably absent from the IMDB set is the Bourne series.

5. FILM SIMILARITY METRICS

Another use of the data collected via GWAP is to create a film similarity metric which would allow us to evaluate how similar films are in the minds of viewers. Determining film similarity, like keyword-extraction, is a problem with no clear programmatic solution. Netflix's dataset, for example, allows them to determine films which are commonly liked together but that is not identical to the problem of determining how similar the films are. The theory which underlies this facet of the project is that films which are often confused on the basis of human-generated keywords probably are similar.

5.1 Computer Generated Similarity

By using the keyword sets generated in the previous section we were able to create film similarity graphs of reasonable quality.

This method was able to extract trivial relationships between films, like those in a series, due to key words which appear many times. The keyword "Batman" is certain to appear across every film in the Batman franchise 5. Unfortunately, outside this method is incapable of clustering films of the same genre or theme together. Films seem to be grouped almost exclusively on the basis of the proper-nouns inside of them. Thus, this method is only an effective sequel detection system.

5.2 Confusion-Based Similarity

If a user confuses one film for another on the basis of some keyword set that implies that that keyword set applies to both films which suggests some level of film-similarity. Creating a graph of commonly confused films reveals higher-order structures.

This method is much better at detecting film similarity. The clusters extend beyond the film-series to genres and even more specific categorizations 6.

One potential concern with this application is that film-confusion does not necessarily imply film-similarity. Additionally, users often are not familiar with the film they are guessing and are more prone to guessing popular films. This introduces a bias where the system believes disproportionately that well-known films are similar to unrelated films. Additionally, films can have elements which would lead a player to confuse them, but the films are actually in radically different genres. This phenomenon is perhaps best exemplified by the fact that the system believes that "Life of Brian" is similar to "The Passion of the Christ". Although both films involve the life of Jesus, the films are otherwise entirely dissimilar in tone and content.

5.3 Film Suggestion

One potential application for a dataset of this sort is a film suggestion program. Instead of taking the Netflix approach of suggesting films to people by looking at past rankings made by other users, the system would recommend films that are deemed to be similar to films that the user has liked in the past. A use-case for this system would be a user that knows he wants to watch an action film similar to some other film he liked in the past. Netflix would only be able to recommend films that others that liked that film liked. That means that Netflix might recommend films by the same director, but not in the action-genre. This system would only recommend films that are "similar" and so would better serve this user's needs.

A major limitation of this system is the number of films included in the game. There are under 1000 films in the dataset. These films were chosen for being the most recognized and modern works and yet we find, anecdotally, that players often have not seen the films. For a film reccomendation system to truly be effective one would need a dataset with a size on the order of Netflix's, many thousands of films. The problem is that the system begins to break down as more films are added. The more obscure the films we present the less useful the incoming data becomes since fewer players will recognize the film. Note that both the suggester and the guesser need to have heard of the film for the system to produce usable results. The game is signficantly less enjoyable when the films are more obscure. While a difficulty level system has been considered there clearly is insufficient tracker to power a suggestion engine of any significant utility.

6. GWAP ENGINEERING

One of the most important and overlooked aspects of GWAP is the engineering required to make the project run properly. When writing software that the public is going to interact with a higher level of quality and robustness is required.

Table 1: Predicted Keywords for *Enemy of the State*

Predicted, in IMDB	Law, surveillance
Valid, not in IMDB	Actor, Apartment, Automobile, Binoculars, Burglary, Cover-up, Criminal, Cruelty, Deceit, Employment Dismissal, FBI Agent, House, Marital Problem, No Title at Beginning, Outer Space, Pay Phone, Pickup Truck, Political Corruption, Tape Recorder, Taxi, Telephone, Tunnel, Villain, Weapon
Invalid, not in IMDB	Altered Studio Logo, Alternative History, Android, Archery, Arson, Artist, Attempted Rape, Attraction, Babe Scientist, Bare Butt, Baseball Bat, Basketball, Bath, Best Friend, Birthday, Bisexual, Black Bra, Black Cop, Blonde, Boat Accident, Bong, Book, Bravery, Bridge, Broken Glass, Broken Leg, Burnt Body, Catholic, Chapter Headings, Cheating, Chess, Child Abuse, Childbirth, Christ Allegory, Confession, Cowboy, Cult Figure, Danger, Dark Hero, Dating, Desert Eagle, Disaster, Drug Addict, Dwarf, Eccentric, Electrocution, Engineer, England, Evil Man, Existentialism, Exploding Helicopter, Eye Gouging, Femme Fatale, Fish, Fistfight, Flamethrower, Flash Forward, Flatulence, Florida, Fondling, Gatling Gun, Golf, Grandmother-grandson relationship, Hanging, Hate, Heist, Horror Movie Remake, Interview, Island, Jeep, Jump Through Window, Jungle, Kissing, Kitchen, Lieutenant, Loneliness, Loss of Mother, Loss of Son, Mad Scientist, Male Bonding, Manhattan New York City, Marriage Proposal, Mayor, Melodrama, Mental Illness, Mercilessness, Mercy, Military Officer, Mission, Mistaken Identity, Mob Hit, Molotov Cocktail, Morgue, Occult, Panties, Part of Trilogy, Pilot, Pizza, Poetry, Poker, Police Brutality, Poverty, Railway Station, Rat, River, Russia, Russian Mafia, Sabotage, Sadism, School, Second Part, Secret, Sex Talk, Slow Motion, Smoking, Space, Space Travel, Spaceship, Surprise After End Credits, Survival, Teen Angst, Terrorist, Third Part, Time Travel, Trust, Underwater, Vandalism, Vietnam, Vulgarity, Widow
Not predicted, in IMDB	African American, Baltimore Maryland, Blackmail, Blood Splatter, Cat, Chase, Christmas, Claustrophobia, Computer, Confrontation, Congressman, Conspiracy, Corruption, Disbelieving Authorities, Distrust of Government, Dog, Echelon, Evidence, Ex-girlfriend, False Accusation, Father-Son Relationship, Frame-Up, Friendship, Fugitive, Gadget, Government Corruption, Helicopter, Innocence, Intelligence, Lethal Injection, Mafia, Marriage, Mexican Standoff, Middle Class, Murder, Neo-noir, NSA, On The Run, Paranoia, Political Thriller, Politics, Privacy, Revenge, Running, Satellite, Secret Hideaway, Shootout, Shot in the Chest, Shot to Death, Suspense, Video Footage, Video Voyeurism, Videotape, Violence, Washington DC, Wiretap, Wisecrack Humor

Table 2: Movies labeled with "surveillance"

Not Labeled by IMDB	*10 Things I Hate About You*, *12*, *2012*, *48 Hrs.*, *Air Force One*, *All the King's Men*, *Analyze That*, *Analyze This*, *Angels & Demons*, *Apollo 13*, *Army of Darkness*, *As Good as It Gets*, *At First Sight*, *Austin Powers: International Man of Mystery*, *Barry Lyndon*, *Blood Simple.*, *Cherry Falls*, *Clueless*, *Contact*, *Cube*, *Deep Cover*, *Edward Scissorhands*, *Escape from New York*, *Finding Nemo*, *Fletch*, *Forrest Gump*, *Frozen River*, *Funny People*, *Gamer*, *Ghost*, *Grand Hotel*, *Groundhog Day*, *Hackers*, *Harold and Kumar Go to White Castle*, *Hotel Rwanda*, *In the Loop*, *Jaws 2*, *Killing Zoe*, *Lake Placid*, *Land of the Dead*, *Liar Liar*, *Magnolia*, *Major League*, *Midnight Express*, *New York Minute*, *Ordinary People*, *Orgy of the Dead*, *Panic Room*, *Pineapple Express*, *Pitch Black*, *Platoon*, *Rachel Getting Married*, *Real Genius*, *Repo Man*, *RocknRolla*, *Saving Private Ryan*, *Scarface*, *Scream 3*, *Shakespeare in Love*, *Sister Act*, *Slither*, *Star Trek: Generations*, *Storytelling*, *Suspect Zero*, *Swingers*, *Taking Lives*, *Tall in the Saddle*, *The Addams Family*, *The Assignment*, *The Battle of Shaker Heights*, *The Birds*, *The Bourne Identity*, *The Bourne Supremacy*, *The Chronicles of Narnia: The Lion, the Witch and the Wardrobe*, *The Crying Game*, *The Curious Case of Benjamin Button*, *The Devil's Advocate*, *The Distinguished Gentleman*, *The French Connection*, *The Italian Job*, *The Limey*, *The Matrix*, *The Nightmare Before Christmas*, *The Proposal*, *The Sweet Hereafter*, *The Talented Mr. Ripley*, *Tin Men*, *Tomorrow Never Dies*, *Twin Peaks*, *Unforgiven*, *Vanilla Sky*, *We Own the Night*, *White Jazz*, *Wild at Heart*
Predicted, labeled by IMDB	*Enemy of the State*, *The Departed*
Not predicted, labeled by IMDB	*8MM*, *Burn After Reading*, *Casino*, *Chinatown*, *Hard to Kill*, *Klute*, *L.A. Confidential*, *Pi*, *Smokin' Aces*, *Strange Days*, *Swordfish*, *The Incredibles*, *The Rocky Horror Picture Show*, *The Salton Sea*, *The World Is Not Enough*, *There's Something About Mary*, *Traffic*, *True Lies*

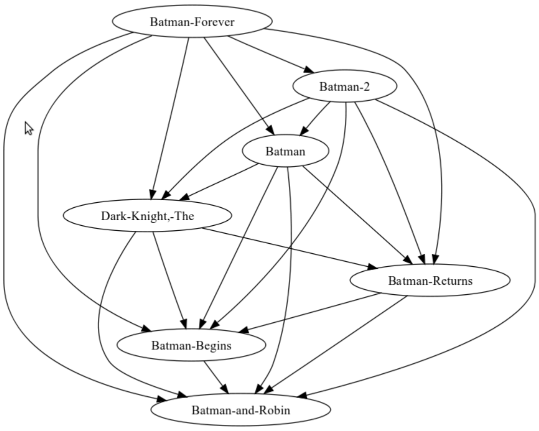


Figure 5: Batman film-series subgraph generated from programatically generated keywords. Edges indicate film similarity.

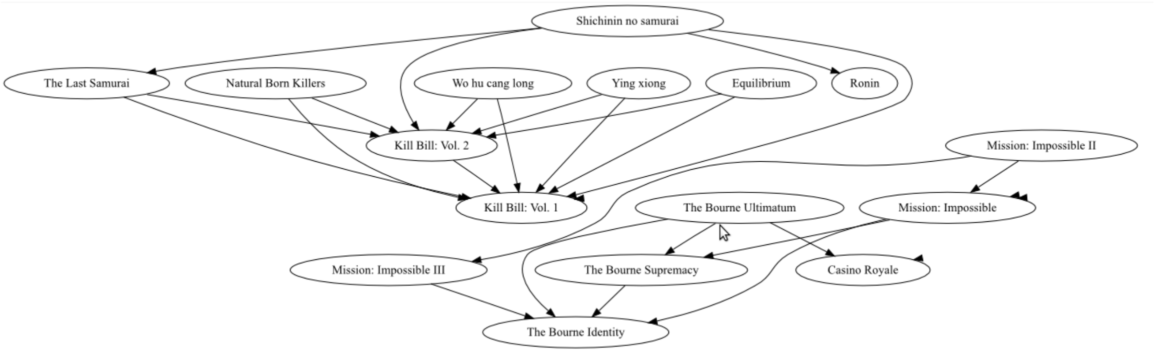


Figure 6: Two subgraphs taken from the user film-confusion dataset. These cluster seems to represent spy films and action-films involving swords.

During the course of the project a great deal more time than anticipated was diverted towards rewriting the user interface and improving the user experience.

6.1 User Experience

Having a good user experience is critical in a GWAP project as the quality of the data that enters the database depends on the contributions of the users. The main user interface is shown in the adjacent figure 7. Users that are unsatisfied with the implementation are less likely to continue to play. The game must be fun and easy to play. This means fixing minor bugs and concerning oneself with minutiae. Features such as a leaderboard were added to entice users to spend more time. Unfortunately, it seems that the user experience of the new version of "Rankmaniac" is lower in quality that the original. Fewer users played the game and traffic numbers failed to spike as they had in the original project. The number of players that signed in and made it to the leaderboard is less than a dozen. This is probably due to the unintuitive interface.

6.2 Google App Engine

When moving from the original Rankmaniac to the new version, we switched from the LAMP stack to Google App Engine. This was done because we believed the distributed nature of Google App Engine's servers would make the game more responsive and functional. Traffic unfortunately never reached numbers which could justify the need for distributing the load. This means that we were left with the disadvantages of Google App Engine with none of the advantages. It is clear that Google App Engine is not intended as a game platform since it severely restricts the length of connections and the number of connections that can be made at any given time.

While App Engine is well-suited for some tasks, data manipulation, data import and data export are all things that it does very poorly. This is the case precisely because of its distributed nature. Unfortunately these sorts of tasks are all crucial for this project. The film listing is not easily distributed across many computers. What's more, Google App Engine suffered a serious outage during the project which resulted in some lost data and downtime. We cannot reccomend Google App Engine for similar future projects.

6.3 GWAP Game Design

For any given problem there are many ways a GWAP can be setup to encourage players to solve the problem. Deciding how the game will be played is a crucial step. Making it too direct can make the game seem like a chore, but making too indirect damages the quality of the data. While we would might get better data if we just asked users to provide keywords for a given film that game-play would not be especially compelling.

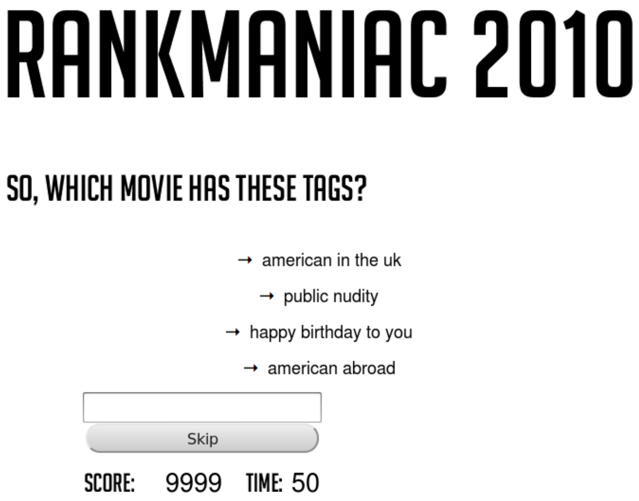
Having the expected input from the user being limited to a small range helps to prevent noise from entering the system. Allowing nearly unfettered contact between the users also creates problems as users can "cheat" and use the keyword submission system to send information other than keywords. GWAP.com gets around this problem in the ESP game by not allowing users to see each other's tags and only ending the round when a match occurred. The matched word is then considered legitimate. This then requires that there be two abusers for malicious keywords to be injected into the system.

The question of how to filter keywords is an important one. Initially the plan was to restrict keywords to words in a dictionary, but it was decided that some proper-nouns were desirable. Similarly, multi-word keywords are part of the IMDB dataset and so we decided to permit them in submissions as well. We used the idea of "taboo" keywords only to prevent the player from directly sending the name of the film (or strings close to the name). This lax policy resulted in submitted tags such as "haven't seen this movie", "haven't seen this film either" and "hellp [sic]". Restricting the set of keywords to strings of dictionray words stops the last of these keywords, but not the first two. Allowing only 1-word keywords stops the first problem, but users could just send their messages one word at a time. Imposing cool-down periods between keywords was considered but discarded because it would make the game frustrating to play. In the end, the only way we believe this sort of abuse could be stopped would be through a peer-review process where players would turn their partner in for not playing by the rules.

Having such a large range over which users are submitting inputs makes many problems. The tag-space becomes very sparse since it is unlikely for users to select the same multi-word keywords. This is especially true when users are not being encouraged to select keywords that others are also likely to select. The more descriptive keywords are also less likely to be used. Good keywords which appear infrequently are unhelpful since they are difficult to distinguish from noise or bad keywords which also appear infrequently. Contributing to the sparseness of the dataset is the fact that most people are not familiar with all of the films in the dataset. Unlike the games at GWAP.com, Rankmaniac requires some prior knowledge which restricts the set of people that can play. It also biases the dataset away from obscure films and towards more popular works.

Malicious injection into the system is an important problem with no obvious solution. In a multiplayer game like this we do not want the players to collude and have a channel of communication outside of the control of the game. However because of the low traffic numbers on Rankmaniac it is trivial for two people to arrange for themselves to playing against one another by entering the matchmaking process at the same time. The users could then proceed to obtain very large scores without contributing any useful information to the project. The only saving grace here is that there is little incentive for players to cheat.

For any given problem there are many ways a GWAP can be setup to encourage players to solve the problem. Deciding how the game will be played is a crucial step. Making it too direct can make the game seem like a chore, but making it too indirect damages the quality of the data. The project rests on the assumption that keywords which aid in guessing a film are also highly descriptive of the film. Unfortunately, there are words and phrases which can bring to mind a film without truly being a keyword. It is unclear whether actors names or character names should be allowed as keywords.



RANKMANIAC 2010

SO, WHICH MOVIE HAS THESE TAGS?

→ american in the uk

→ public nudity

→ happy birthday to you

→ american abroad

SCORE: 9999 TIME: 50

Figure 7: The main user interface for the game is left very simple.

Ultimately it seems that there is no objective measure of the quality of a keyword outside of the game's assumptions. Perhaps the problem then is that GWAP functions best in situations where the problem trying to be solved is very clear cut. Reading is a very well defined problem and so the reCAPTCHA system works very well..

6.4 Traffic

From the start of the Rankmaniac website, we used Google Analytics to monitor traffic to the website. A chart of the traffic to the site is shown in Figure 8. Major traffic events have been labeled with their source. College Humor contributed to the first such spike. Stumble Upon delivered periodic bursts of traffic. The Google Adwords campaign near the end of second term for CS144 coincided with the last significant burst of traffic from Stumble Upon. The traffic numbers stayed steady and unfortunately, low, throughout the rest of the experiment. Most of the incoming traffic was still directed via Stumble Upon.

7. CONCLUSIONS

One goal of this project was to build a useful dataset for other projects in the areas of keyword extraction and film analysis. Organizations such as IMDB or Netflix could apply this dataset in order to provide more relevant keywords for films. The keywords which were most "helpful" to identifying movies could be useful in resolving search queries. Additionally, confusion between movies can be used for gauging the similarity of films. As shown in Figure 5, films on a related topic, Batman, are frequently confused within the Rankmaniac dataset. Figure 6 shows how the dataset extends beyond clustering movies centered around the same character; amongst spy films, Mission Impossible is confused with the Bourne Series.

Another goal of the project was to design a program which can efficiently extract meaningful keywords from film scripts. User-generated content was extremely sparse for the task at hand. Users provided poor responses for films which they were not familiar. Unlike with GWAP's ESP, the targets for identification are films, not commonplace things. When applied to the script set using the IMDB dataset for labels, naïve Bayes was too eager in assigning keywords to movies. As examination of the results produced by this algorithm showed, keyword extraction is a far cry from being truly automated. The state of the art in automatic keyword selection vastly underwhelms Vannevar Bush's vision of the future.

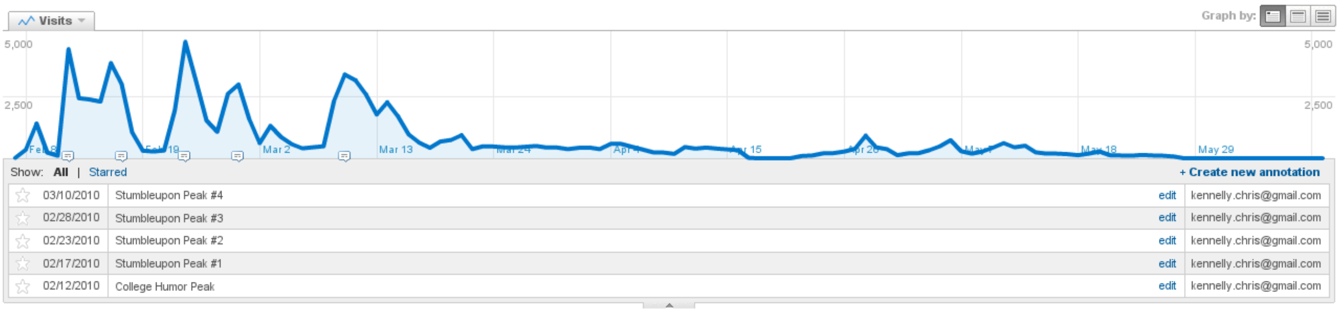
Nonetheless, the Rankmaniac game has provided a useful framework for analyzing the relationships of movies to various tags.

8. FUTURE WORK

Other scientific projects could use this dataset to help determine the most appropriate keywords for films based upon the script or some other features related to the films. For example, these keywords would be the objective output of the function which is attempting to describe the films based upon some automatically obtainable data such as script, the movie transcript, or other data. We could publish this dataset online and it would be useful in other projects concerning films, projects like the Netflix prize.

We did not exhaust all possible ways to analyse the data that we are already collecting through the game. For example, we ignore the order in which the tags arrive and the guesses are made. This data is important because when a user guesses they are making a statement about the keywords they have seen so far. This information could increase the predictive power of our models. Additionally, as previously mentioned, an adaptive difficulty setting would allow us to introduce more films into the game making the data more practically useful.

The method in which the game operates also could be changed



☆	03/10/2010	Stumbleupon Peak #4	edit	kennelly.chris@gmail.com
☆	02/28/2010	Stumbleupon Peak #3	edit	kennelly.chris@gmail.com
☆	02/23/2010	Stumbleupon Peak #2	edit	kennelly.chris@gmail.com
☆	02/17/2010	Stumbleupon Peak #1	edit	kennelly.chris@gmail.com
☆	02/12/2010	College Humor Peak	edit	kennelly.chris@gmail.com

Figure 8: Google Analytics-monitored Visitor Traffic Per Day

to improve the quality of our data. By limiting the users' ability to inject noise and spurrious keywords we would increase the fraction of desirable data. For example, the system could prevent the users from submitting keywords except on an approved list. Or, the game could be changed entirely such that both players name films similar to the film they have been shown. This way the dimension of the user supplied input is small and so the dataset becomes more dense. Alternatively, we could couple the automatic keyword generator with the game by supplying one user with a list of a few keyword choices which they then select from.

9. REFERENCES

[1] V. Bush and J. Wang. As we may think. *Atlantic Monthly*, 176:101–108, 1945.

[2] J. Gray. What next?: A dozen information-technology research goals. *J. ACM*, 50(1):41–57, 2003.

[3] L. von Ahn. Games with a purpose. *IEEE Computer Magazine*, pages 96–98, 2006.

[4] L. von Ahn. Designing games with a purpose. *Communications of the ACM*, pages 58–67, 2008.