**Module 11**

**Pitfalls of applying AI to Information Retrieval tasks**

**Outline**

• Introduction

• Information Retrieval

• Artificial Intelligence for Information Retrieval

• Disadvantages of AI techniques in Information Retrieval

• Challenges

**1. Introduction**

This module intends to offer the learners the opportunity to deepen their understanding of the relationship between Information retrieval (IR), and Artificial Intelligence (AI), with an emphasis on investigating at several applications of AI to IR in different use cases, e.g. web search, semantic web search, digital library search. Both IR and AI fields developed in parallel during the early days of computers. The fields of AI and IR share a common interest in developing more capable computer systems. Integration of AI and IR lead to develop methods to learn user's information needs and extract information based on what has been learned, and represent the semantics of information.

This module describes the most prominent approaches for applying AI technologies to IR. This module can be integrated with CS 6821 Information Retrieval course.

**2. Information Retrieval**

The amount of available information is growing at an incredible rate, for example the Internet and World Wide Web. Information stored in many forms such as images, text, video, and audio. IR is a way to separate relevant data from irrelevant. IR field has developed successful methods to deal effectively with huge amounts of information.

Information retrieval systems provide access to collections of thousands, or millions, of documents, such that users can recover any one by providing an appropriate description. Typically, users iteratively refine the descriptions they provide to satisfy their needs, and retrieval systems can utilize user feedback on selected documents to indicate the accuracy of the description at any stage [1].

This module focuses on the relationship between AI and IR, however if learners do not take CS 6821 they can see all resources which are provided in the reference section [2, 3, 4]. The resources will give the learners some knowledge about different types of IR models, the components involved, and the techniques used in IR to understand the mechanism behind search engines displaying results.

To recap, IR comprises of four key elements, Document Representation (D), Query Representation (Q), A framework to match and establish a relationship between D and Q. A ranking function that determines the similarity between the query and the document to display relevant information. The most-adopted similarity-based classical IR models Boolean, Vector Space and Probabilistic models. As IR needs to deal with vague knowledge, exact processing methods are not appropriate. Vague retrieval models like the probabilistic model are more suitable [5].

From an AI point of view, we can discuss the specific properties of IR with respect to AI’s concerns with the nature of knowledge representation, reasoning, and learning. AI might profitably learn from all forms of information management.

**3. Artificial Intelligence for Information Retrieval**

AI methods are employed throughout the standard IR process and in fact IR is an AI task. Artificial intelligence methods are employed throughout the standard information retrieval process. Several approaches use machine learning for optimizing of the matching between term and document. Neural networks for document clustering & preprocessing have been applied widely in IR, so learning from users has been an important strategy to improve systems. Therefore, Machine Learning (ML) methods have been used to improve the user interface.

An intelligent IR system can simulate the human thinking process on information processing and achieve information and knowledge storage, retrieval and reasoning, and to provide intelligence decision. According to Bates [16], levels of user and system involvement can be:

1. No system involvement (User comes up with a tactic, formulating a query, coming up with a strategy and thinking about the outcome)

2. User can ask for information about searching (System suggests tactics that can be used to formulate queries e.g. help)

3. User simply enters a query, suggests what needs to be done, and the system executes the query to return results.

4. First signs of AI. System actually starts suggesting improvements to user.

5. Full Automation. User queries are entered and the rest is done by the system.

Like many other science domains, AI methods currently used in IR Systems. For example in information extraction like Web Crawlers, for information integration like Mediator Techniques, and Semantic Networks

….

**4. Disadvantages of AI techniques in Information Retrieval**

IR can not take the Human Factor completely out of the equation. A program which takes a query as input, and returns documents as output, without affording the opportunity for judgment, modification and especially interaction with text, or with the program, is one which would not qualify as an IR system at all.

On the other hand an IR systems should not ignore user interaction and relevance feedback. Some processes which can not be performed by any other component than the user. “Judgment” is a process which can only be performed by the user [9]. Further obstacles are in AI/IR Systems such as Common Sense Reasoning, Natural Language Processing, Knowledge Acquisition, Representation, and Maintenance, Level of Effort, Technical Expertise, and Expense.

When AI is combined with conventional and other innovative computing tools, it can be a powerful tool. But, it is not an easy task to master those tools and employ them skillfully to build truly significant intelligent systems. AI models like the deep learning will probably not replace traditional IR approaches soon. However, the application of machine leaning models can make an IR system more powerful [].

**5. Challenges**

**Recommender Systems**

A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that provide suggestions for items that are most pertinent to a particular user. Typically, the suggestions refer to various decision-making processes, such as what product to purchase, what music to listen to, or what online news to read. Recommender systems are particularly useful when an individual needs to choose an item from a potentially overwhelming number of items that a service may offer. The goal of the recommendation system is to realize the artificial intelligence of user processing information. In recent years, deep learning technology has been an important part of data mining and artificial intelligence research. Deep learning can use a complex multivariate function as a training target by means of a grid environment. Grid resources center allocates the calculated tasks to the idle computing resources in the platform. After the computing resources obtain the scheduling tasks, the data processing begins. After the data processing is completed, the computing resources will be released, and then the information server will feed back the running status of all computing resources in the grid to the grid resources center.

There are two basic approaches to recommending:

**Collaborative Filtering** – Keep track of how many users have rated a wide range of items in a database. Find users who are similar to a given user and whose ratings are highly correlated with the user's own. recommend products that were highly rated by these users who are similar to you but were not rated by you. This strategy is used by nearly all current commercial recommenders.

**Content-based Filtering** - Content-based filtering is another strategy that is frequently used when constructing recommender systems. A description of the item and a profile of the user's interests serve as the foundation for content-based filtering techniques. These techniques work best when information about the item (name, location, description, etc.) but not the user is known. Content-based recommenders approach recommendations as a user-specific classification issue and develop a classifier for a user's preferences based on the characteristics of an item.

**The Netflix prices**

The Netflix Prize was one of the occasions that sparked research in recommender systems. In a competition that lasted from 2006 to 2009, Netflix offered a $1,000,000 top prize to the team that could use a dataset of more than 100 million movie reviews to provide suggestions that were 10% more accurate than those provided by the business's current recommender system. This competition sparked a renewed interest in developing brand-new, more precise algorithms. BellKor's Pragmatic Chaos team was awarded the $1,000,000 main prize on September 21st, 2009 following tie-breaking procedures. The most accurate algorithm in 2007 combined 107 independent computational techniques into a single forecast using an ensemble method. The winners, Bell et al., declared the following:

**Performance measures**

Evaluation is crucial in determining how well recommendation systems work. Three forms of evaluations—user studies, online evaluations (A/B testing), and offline evaluations—can be used to assess the performance of recommender systems and compare various strategies.

The root mean squared error and mean squared error are the two most often used metrics; the latter was used to the Netflix Prize. A recommendation method's quality may be evaluated using information retrieval measures like recall and accuracy or DCG. Aspects like diversity, innovation, and coverage are also seen to be crucial in assessment. However, several of the traditional evaluation methods have received harsh criticism.

**Activity:**

Try making a recommended system by following the steps in <https://365datascience.com/tutorials/how-to-build-recommendation-system-in-python/>

You can use any other resource you like to make one.

“AM”

**TERM WEIGHTING**

BY VAISHNAVI ANKAM

Index term weights reflect the relative importance of words in documents and are used in computing scores for ranking. The specific form of a weight is deter- mined by the retrieval model. The weighting component calculates weights using the document statistics and stores them in lookup tables. Weights could be calculated as part of the query process, and some types of weights require information about the query, but by doing as much calculation as possible during the indexing process, the efficiency of the query process will be improved.

One of the most common types used in older retrieval models is known as TF\_IDF weighting. There are many variations of these weights, but they are all based on a combination of the frequency or count of index term occurrences in a document (the term frequency, or TF) and the frequency of index term occurrence over the entire collection of documents (inverse document frequency, or IDF). The IDF weight is called inverse document frequency because it gives high weights to terms that occur in very few documents. A typical formula for IDF is log N /n, where N is the

**Approaches to implement term weighting using pyterrier:**

· PL2

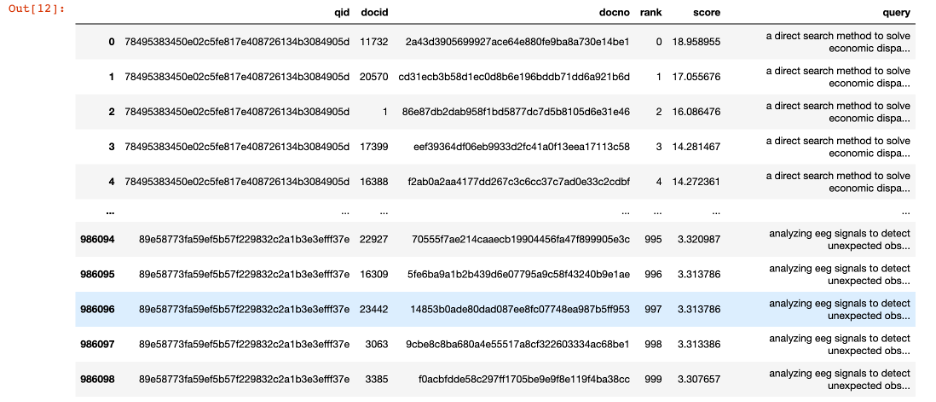
· BM25

· TF\_IDF

**Activity**:

[**https://pyterrier.readthedocs.io/en/latest/terrier-retrieval.html**](https://pyterrier.readthedocs.io/en/latest/terrier-retrieval.html)

**SAMPLE OUTPUT for irds: beir/scidocs dataset:**

****

**EVALUATION CRITERIA:**

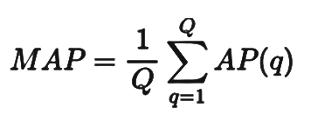
· Mean Average Precision (MAP)

· Normalized Discounted Cumulative Gain (NDCG)

**MEAN AVERAGE PRECISION:**

The basic formulation behind MAP evaluation is that for any given query q, its corresponding average precision AP is calculated, and then the mean of the all these AP scores evaluates as the MAP score, which quantifies how good our model is performing for that query. If we want to evaluate average precision across multiple queries, we can use the MAP.

The score calculation can be done using this formula:



where,

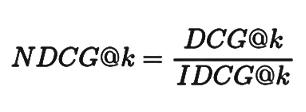
Q - total number of queries, and AP(q) – average precision of the query q

**NORMALIZED DISCOUNTED CUMULATIVE GAIN:**

The goal of NDCG is similar to that of the MAP evaluation metric. However, NDCG further tunes the recommended lists evaluation. It also takes into account the ordering of results. i.e. Highly relevant items should come before medium relevant items, which should come before non-relevant items.

To allow a comparison of DCG across queries, we can use NDCG that normalizes the DCG values using the ideal order of the relevant items. Thus, we get NDCG scores with a range between 0 and 1. A perfect ranking would get a score of 1.

The value of the NDCG metric can be calculated as follows:



**Boolean Model in Information Retrieval**

By Varshini Subramaniam Sankaran

The Boolean model of information retrieval, uses *matching technique* to match documents to a user query. Documents are retrieved by checking whether the documents contain the terms in the query. Hence a document is categorized as relevant or irrelevant and strictly no partial matching. With the use of *Boolean algebra,* words are logically combined with the Boolean operators AND, OR, and NOT.

**Advantages:**

· Easy to implement

· Suppose the result document set is either too small or too big, it clearly tells which query would have produced the bigger or smaller set respectively.

**Drawbacks:**

· *Synonymy: It happens* when many words have the same synonym. Eg: "car" and "four-wheeler".

· Polysemy: It happens when one word has many meanings. Eg: If the user query is “bank” it is unclear what it refers to whether is it a financial center, or river bank or a slope on a hill.

**Activity:**

Consider 6 documents with a vocabulary of 7 terms

· Doc1 = ‘Cat Dog Rat’

· Doc2 = ‘Cat Rat Sheep’

· Doc3 = ‘Lion Tiger’

· Doc4 = ‘Cat Dog Fish Lion Rat Sheep Tiger’

· Doc5 = ‘Cat Dog Rat Sheep’

· Doc6 = ‘Lion’

Based on the Boolean Model, find the documents that contain the terms:

1. Cat AND Rat AND NOT Dog

2. Lion OR Tiger

3. Fish AND Sheep

4. NOT Sheep

5. NOT Lion OR NOT Dog

**Cross Lingual information retrieval**

BY Jinalben Luvani

A large amount of information in the form of text, audio, video and other documents is available on the web. These documents ought to provide users with relevant information. The process of finding relevant documents and information within a dataset, such as the World Wide Web(WWW), is referred to as “Information Retrieval”(IR). A web search engine is an IR system that is designed to look for information on the World Wide Web. Information retrieval involves a number of different components. The following are the components of an IR system:

· Crawling: data extracted from the web are fetched and stored.

· Indexing: An index of the fetched documents is created.

· Query: Input from the user

· Ranking: Organization of data according to their relevance to the query

Cross Lingual Information Retrieval(CLIR) systema are in high demand due to the widespread use of the internet and the numerous networked information sources that are available. Cross Lingual information retrieval is used to find relevant information when a user query is written in a different language than the document collection. This allows users to search documents in multiple languages and find relevant information in a way that is useful to them, even when they have little or no linguistic competence in the target language.

**APPROACHES TO CLIR:**

1) Query translation approach

This method involves translating the query into the language of the document. Many different translation strategies, such as dictionary-based translation or more complex machine translations, may be possible. This method's main presumption is that the user is able to read and comprehend documents written in the target language. It is not necessary to have these tools for all language pairs.

2) Document translation approach

With this method, there is no need to translate documents from other languages into the language of the query. Although there are too many documents to translate and each document is fairly huge in comparison to a query, this strategy has scaling problems. Due to this, the strategy is essentially inappropriate.

3) Interlingua based approach

In this instance, a common Interlingua is used to translate both the documents and the query (like UNL). Since the translation must be done online, this method typically requires significant resources.

**ACTIVITY:**

Write down the challenges of CLIR and build the CLIR module.

**References**

1. APPLICATIONS OF MACHINE LEARNING IN INFORMATION RETRIEVAL, Sally Jo Cunningham, James Littin and Ian H. Witten

2. Introduction to Information Retrieval, By Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze

3. Information Retrieval Systems, <https://www.sciencedirect.com/topics/computer-science/information-retrieval-systems>

4. Information Retrieval- A Brief Overview, <https://medium.com/@soumya.vkshukla/information-retrieval-a-brief-overview-173bba8fe0e9>

5. <https://www.researchgate.net/publication/314457405_Artificial_Intelligence_for_Information_Retrieval>

6. Understanding User Goals in Web Search Daniel E. Rose, Danny Levinson, <http://facweb.cs.depaul.edu/mobasher/classes/csc575/papers/www04--rose.pdf>

7. N.J. Belkin, H.M.Brooks and P.J. Daniels, “Knowledge acquisition using discourse analysis”. International Journal of Man-Machine Studies, 27: 127-144, 1987.

8. P. Maes, “Agents that Reduce Work and Information Overload”. Communications of the ACM, 37(7): 31-40, 1994.

9. M. Bates, “Where should the person stop and the information search interface start?” Information Processing and Management, 26(5): 575-59, 1990.

10. K.S. Jones, “Information retrieval and artificial intelligence”, August 1998.

11. Thomas Mandl, “Artificial Intelligence for Information Retrieval”, 2009.

12. D. Fensel & M. Lalmas “Seeking Information: Methods from Information Retrieval and Artificial Intelligence”, 2000

13. Nicholas J. Belkin, “Intelligent Information Retrieval: Whose Intelligence?”

14. C.W. Bailey, “Intelligent Library Systems: Artificial Intelligence Technology and Library Automation Systems”, 1991.

15. J. Cook, A.D. Whittaker, R.H. Thieme, O.R. Smith, and G.Salvendy, “Human Intelligence Models and Their Implications for Expert System Structure and Research”, 1988.

16. M. Bates, “Where should the person stop and the information search interface start?” Information Processing and Management, 26(5): 575-59, 1990.

17. APPLICATIONS OF MACHINE LEARNING IN INFORMATION RETRIEVAL,<https://www.cs.waikato.ac.nz/~ihw/papers/00-SJC-JL-IHW-Applicationml.pdf>

18. Recommender system <https://en.wikipedia.org/wiki/Recommender_system> & <https://www.cs.utexas.edu/~mooney/ir-course/slides_pdfs/Recommenders%20notes.pdf>

19. Term weighting [**https://pyterrier.readthedocs.io/en/latest/terrier-retrieval.html**](https://pyterrier.readthedocs.io/en/latest/terrier-retrieval.html)

***PIR (Private Information Retrieval )***

***By: Srijani Basu***

***1. Necessity of PIR:***

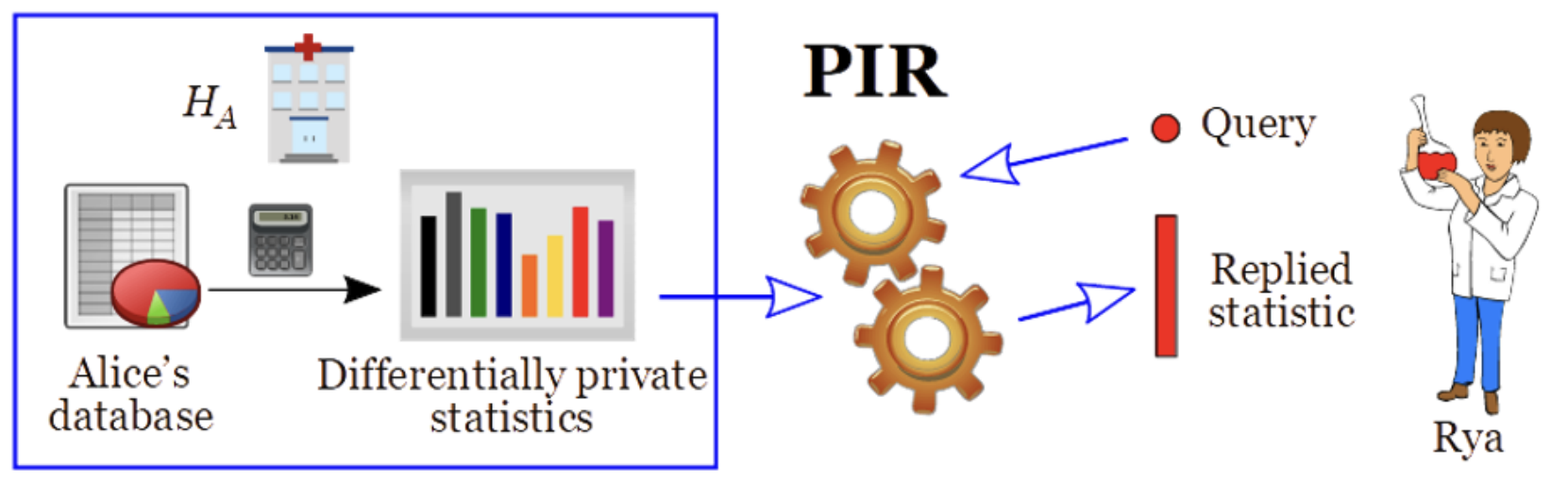
Many recent research projects have focused on how to safeguard a user's personal data while they engage with services. These investigations can be divided into four categories. One of them is PIR-based methods.

***2. PIR (Private Information Retrieval):***

A client can receive a database item from an untrusted server using the PIR protocol without disclosing which item is fetched from the host . Although PIR can be used to construct a query result set in private, it is difficult to avoid a linear private scan of the entire object space.

***2.1. How PIR based approaches work:***

In its most abstract form, private information retrieval lets a user to query an item from a database without revealing the query to the server that is (possibly) storing the data. Suppose Rya is interested in retrieving the ith bit from a database D that belongs to Alice and has n bits in it. Bob can use PIRprotocols to secretly retrieve D[i] without telling Alice what it is. This PIR definition presents a theoretical viewpoint.

****

***3. Reasons Why PIR is challenging :***

i. It is impossible to trust the server that holds the knowledge about the object to process queries and decide what information should be delivered as responses.

ii. As an alternative, transferring this knowledge to the users would necessitate expensive client-side query processing.

***1. Types of PIR Approaches:***

Depending on whether they offer information **theoretic** or **computational** privacy, PIR methodologies can be classified. While the security of computational procedures depends on the intractability of a computationally challenging mathematical problem, information theoretic approaches ensure absolute privacy.

***4.1 "The Computational PIR Approach" Over "The Theoretic PIR Approach":***

The theoretical PIR Approach group's absolute privacy comes at an unaffordable price. For a database of n bits, it has been demonstrated through research that such techniques have an ohm(n) communication cost. As a result, data-intensive and practical applications cannot effectively include information theoretic PIR. In contrast, the computational PIR approaches significantly reduce complexity by assuming some constraints on the server's computational power.

The proposed PIR protocols are still expensive and need a lot of server resources, even though computational PIR has more realistic prices for retrieving items. In other words, even if they can reduce communication complexity, the server must still process every database record. Some researchers even contend that it is much more expensive to discreetly get database objects from the server than it is to deliver the full database to the client.

***4.2. Introducing Hard-ware based PIR Approach:***

A new class of Hardware-based PIR approaches, which put the trust in a tamper-resistant hardware device, has recently come out as a way to get perfect privacy without the high cost of the other methods. When it comes to privacy, these methods rely on a hardware device. This is because the computations they use are very fast. By putting a trusted module close to an untrusted host, these techniques can achieve optimal computation and communication costs compared to the computational PIR approaches.

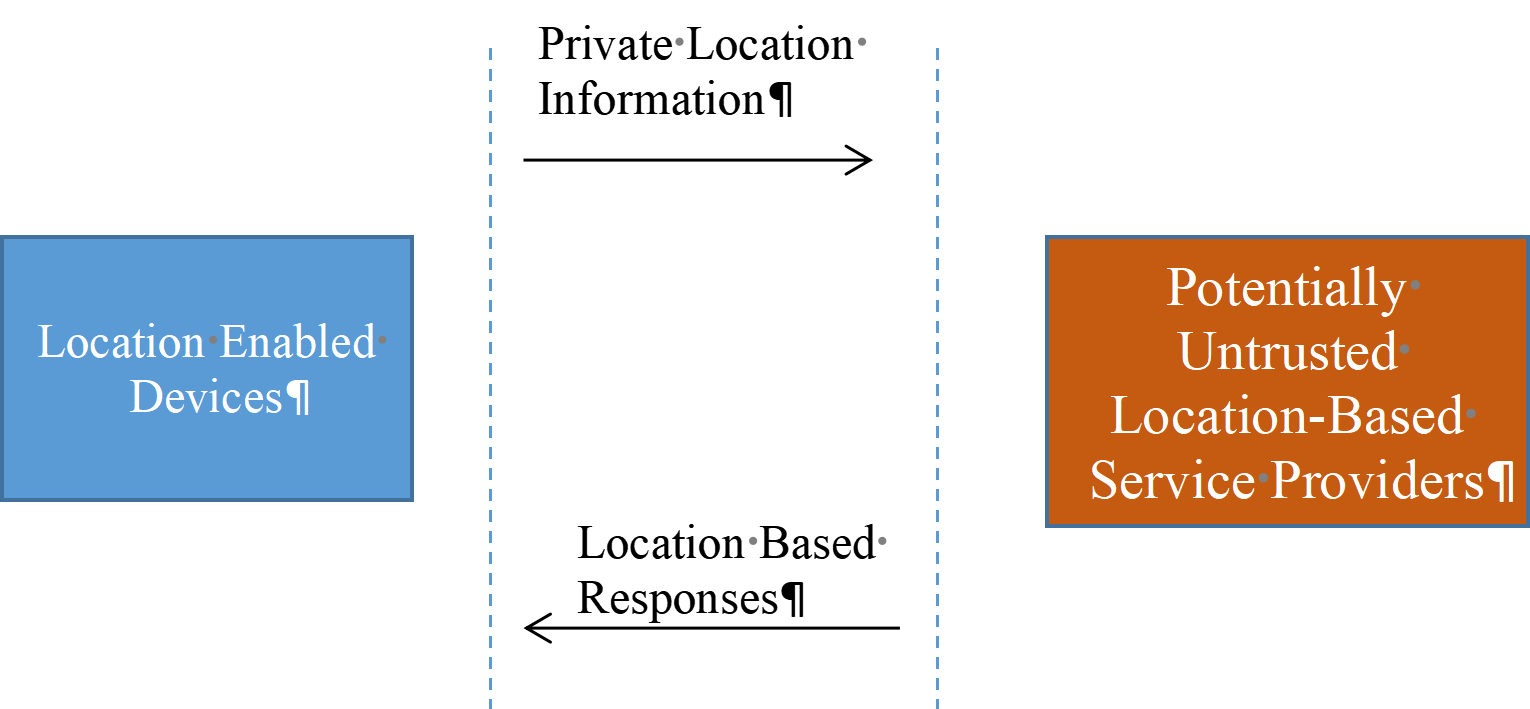
***4.3. Secure Coprocessor:***

A Secure Coprocessor (SC) is a general-purpose computer made to meet strict security requirements that make sure the code on it runs without being seen or changed, even if an attacker is in the same room. Hardware cryptographic accelerators are built into these devices. They make it easy to use encryption algorithms like DES and RSA. Secure coprocessors have been used successfully in the real world for things like data mining and trusted co-servers for Apache web-server security when the server hosting the data is not trusted. The idea behind using a secure coprocessor for PIR operations is to put a trusted entity as close as possible to the untrusted host to hide the selection of desired records inside a black box.

***4.4. Secure Coprocessor Vs Location Server:***

A location server consists of a number of programs using shared memory, whereas a secure coprocessor is a hardware device specifically developed to execute a specified task in addition to being created as a tamper-resistant device. Second, employing a location server necessitates users' full trust, not just that of the designer, unlike using a secure coprocessor, where users need only have faith in the designer.

Last but not least, the secure coprocessor in our environment is primarily a computer that receives its required information, per session, from the server as opposed to a server that both maintains location information and handles spatial requests.



***5. Real World Application Of PIR:***

i. PIR schemes to make the Tor anonymous communication network scale better and examine the resulting design trade-offs.

ii. Tailor and extend the cryptography of PIR to solve the real-world problem of preserving access privacy in an electronic commerce transaction.

***References:***

I. Privacy in Location-Based Applications, Lecture Notes in Computer Science, Volume 5599. ISBN 978-3-642-03510-4. Springer Berlin Heidelberg, 2009, p. 59

II. Femi George Olumofin (2011). Practical Private Information Retrieval. UWSpace.<http://hdl.handle.net/10012/6142>

III. <http://course.ece.cmu.edu/~ece733/lectures/20-pir.pdf>

IV. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: MobiSys'03, San Francisco, CA

V. Bettini, C., Wang, X.S., Jajodia, S.: Protecting privacy against location-based personal identification. In: SDM'05, Trondheim, Norway 185-199

**Probabilistic Models**

**By: Anwesh Gaurav**

It is a retrieval model. In the probabilistic model, the ranking system is based on probability. This model ranks documents in decreasing probability of usefulness.

Usefulness is nothing but relevance. A document is is retrieved if it is relevant.

We need to know Bayes’ Theorem for understanding the above inequality which is given below:

P(H/E)= P(H)P(E/H) / P(E)

P(H/E) is the probability of the hypothesis given evidence.

**Probability Ranking Principle (PRP):**

Using a probabilistic model, the obvious order in which to present documents to the user is to rank documents by their estimated probability of relevance with respect to the information need: P(R = 1|d, q). This is the basis of the Probability Ranking Principle (PRP):

“If a reference retrieval system’s response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.”

x = document in the collection.

R= relevance of a document

NR = non-relevance

P(R/x) = Probability that a retrieved document x is relevant.

As per Bayes Theorem:

P(R/x)= p(x/R).p(R)/p(x)

p(NR/x) = p(x/NR). p(NR)/p(x)

P(R) = probability of retrieving a relevant document

P(x/R) = probability that if a relevant document is retrieved it is x.

The decision rule is that for document to be relevant following condition needs to be true:

p(R/x) > p(NR/x)

I.e. If the above condition is true, the given document x is relevant otherwise it's not.

The assumption made in PRP is that the different words in a query are independent which is not quite right. But, it doesn’t affect the mathematics involved. Calculating the exact value of these probabilities can be difficult so we use estimates. For this we have the Binary Independence Retrieval or Binary Independence Model. It makes two assumptions:

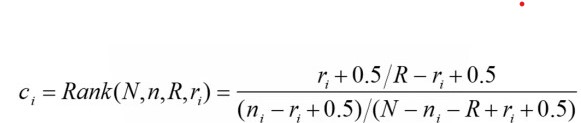
1. Terms in a document are independent of each other.
2. Different documents can be modeled as the same vector.

**The Binary Independence Model (BIM)** is the model that has traditionally been used with the PRP. It introduces some simple assumptions, which make estimating the probability function P(R|d, q) practical. Here, “binary” is equivalent to Boolean: documents and queries are both represented as binary term incidence vectors. That is, a document d is represented by the vector ~x = (x1 , . . . , xM) where xt = 1 if term t is present in document d and xt = 0 if t is not present in d. With this representation, many possible documents have the same vector representation. Similarly, we represent q by the incidence vector ~q (the distinction between q and ~q is less central since commonly q is in the form of a set of words). “Independence” means that terms are modeled as occurring in documents independently. The model recognizes no association between terms. This assumption is far from correct, but it nevertheless often gives satisfactory results in practice; it is the “naive” assumption of Naive Bayes models. Indeed, the Binary Independence Model is exactly the same as the multivariate Bernoulli Naive Bayes model.

Retrieval Status Value:

The aim of information retrieval systems is to retrieve documents that are relevant to the user queries. To reach this goal, they attribute a value to each candidate document; afterwards, they rank documents in the reverse order of this value. This value called the Retrieval Status Value (RSV) is given as:

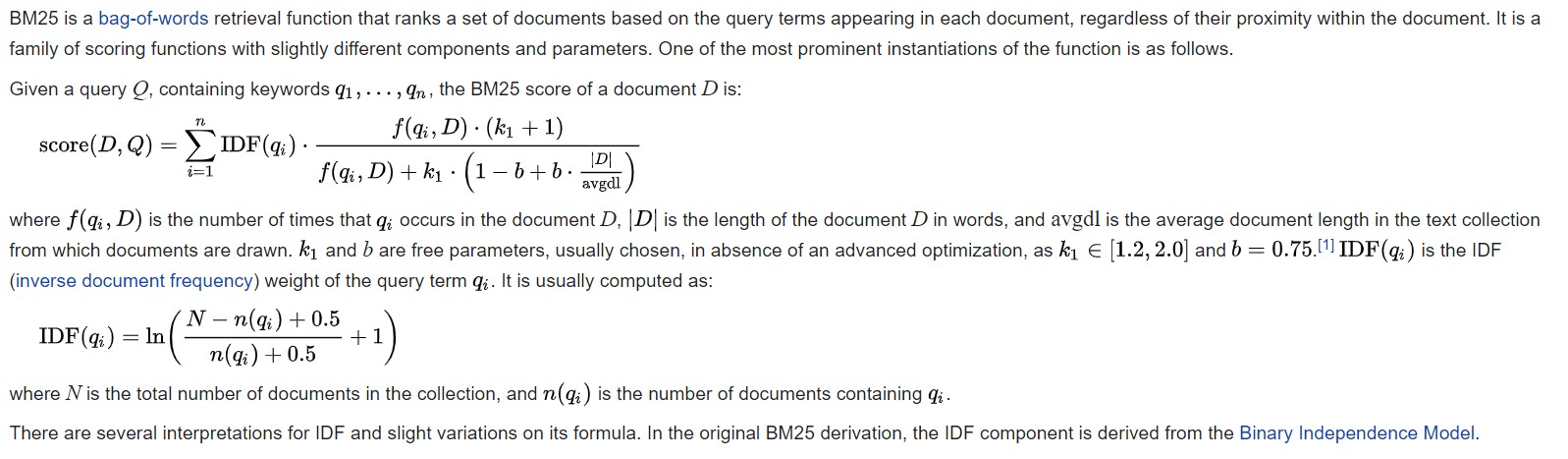


Ranking Formula in Binary Independence Model:

Taking log it becomes:

∑log Ci = log (N-ni  ) /ni

BM25:



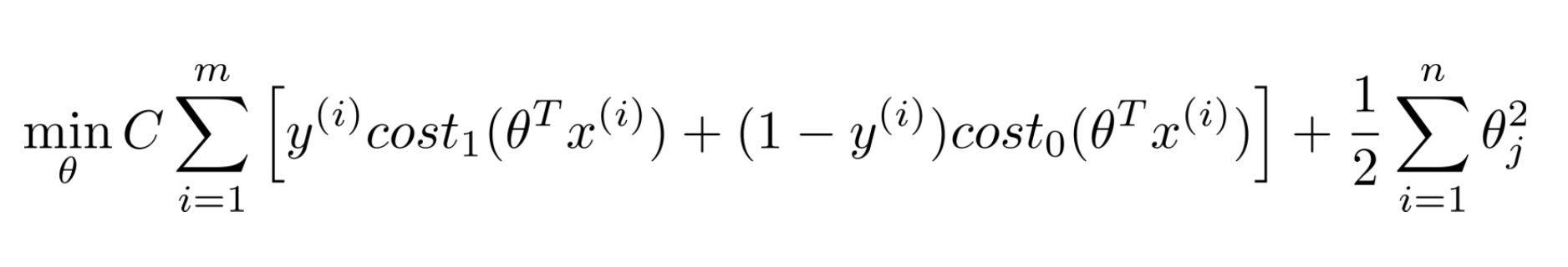
References:

1. Introduction to Information Retrieval, Manning, Raghavan, Schutze
2. Wikipedia

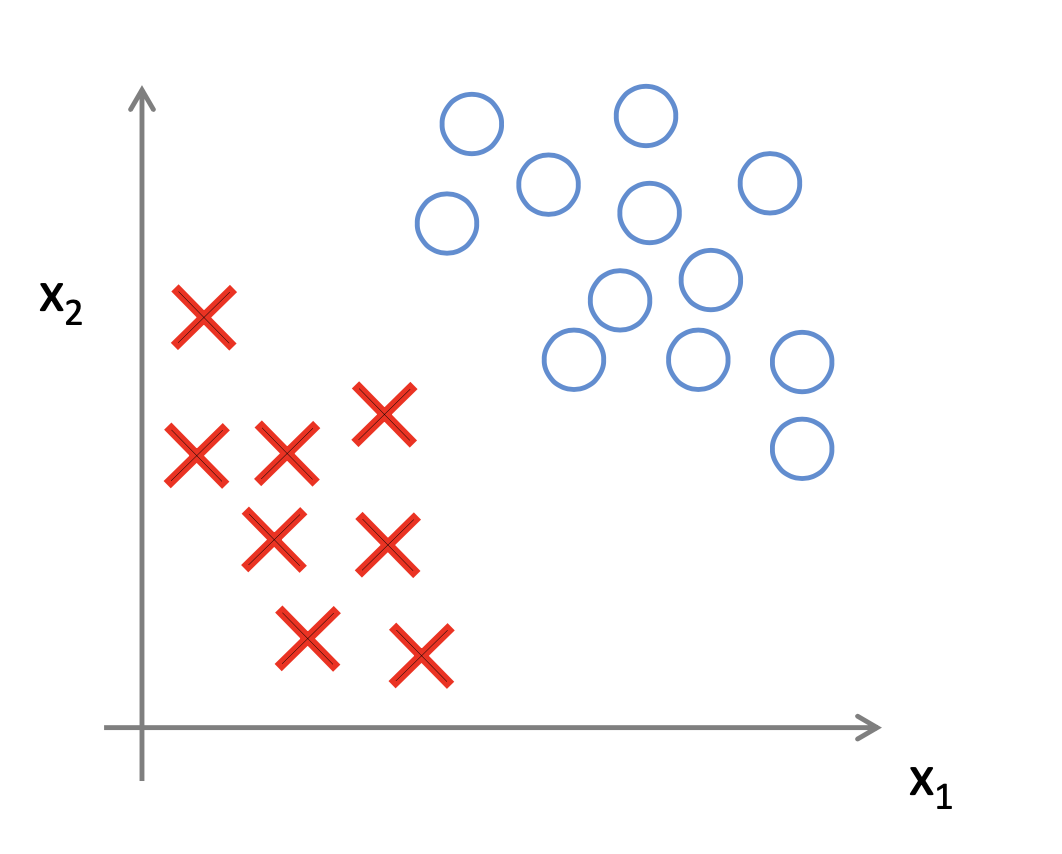
**Support Vector Machines (SVM)**

( Joseph George )

**Support Vector Machine:**

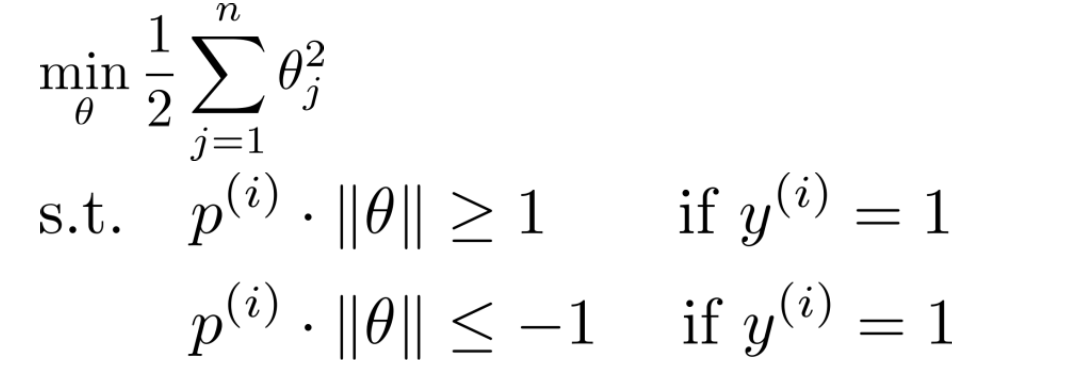


**SVM Decision Boundary : Linearly Separable Case**



Large margin Classifier

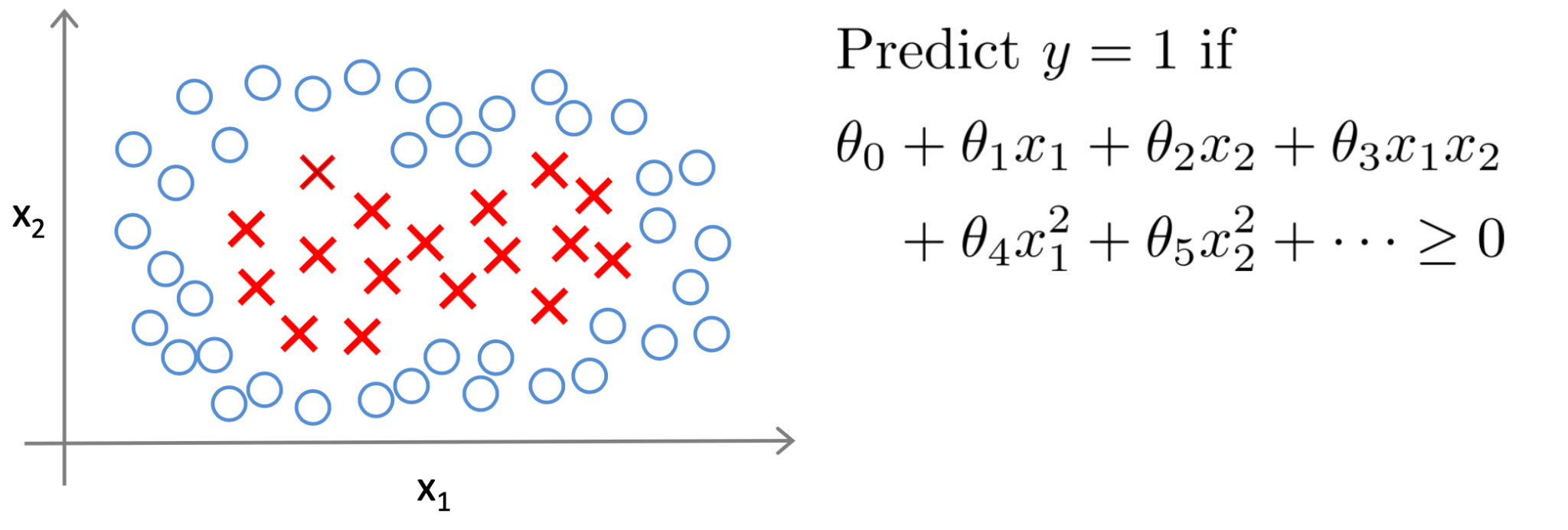
**SVM Decision Boundary**



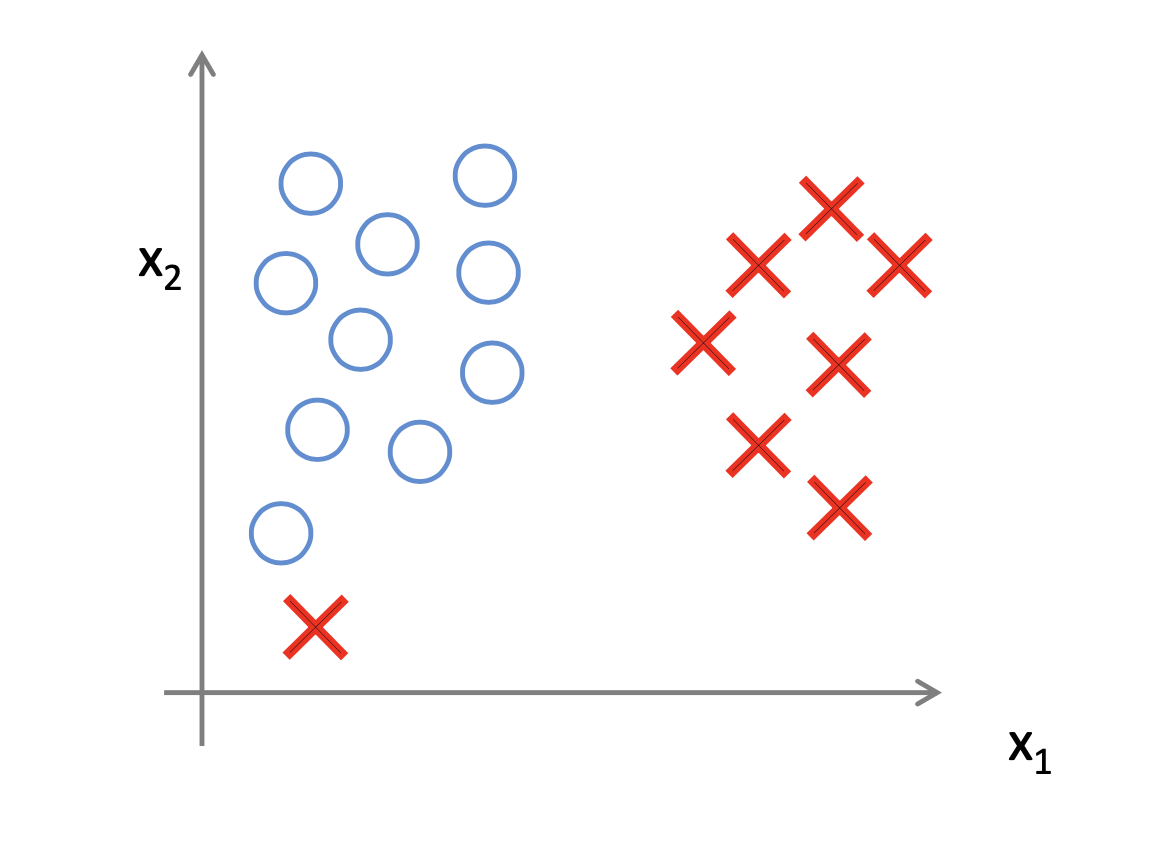
where p^(i) is the projection of x^(i) on to vector . Simplification :  0 = 0



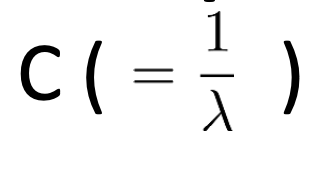
**Non - linear Decision Boundary**



**Large margin classifiers in presence of outliers**

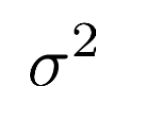


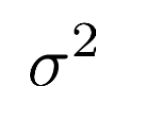
**SVM parameters:**



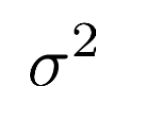
Large C means lower bias, high variance.

Small C means higher bias, low variance.



Large features f(i) very smoothly.

Higher bias, low variance.

Small features f(i) vary more smoothly.

Higher bias, lower variance.

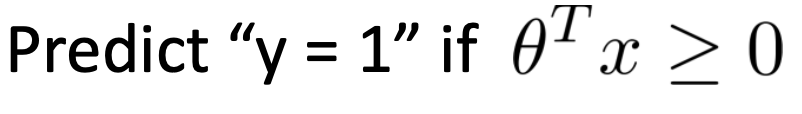
**Using an SVM software package -**

Use SVM packages like liblinear, libsvm to solve for parameters .

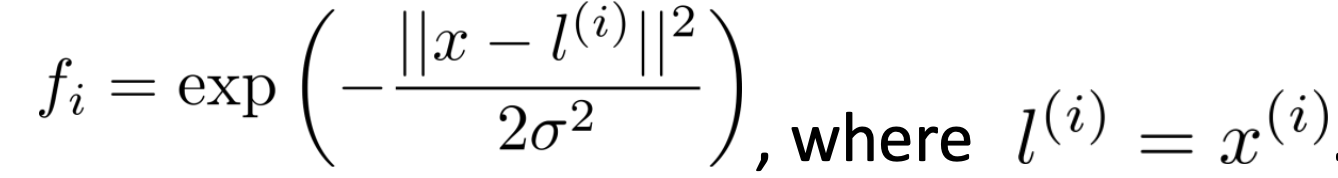
Need to specify:

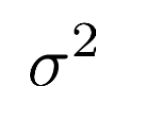
1. Choice of parameter C.
2. Choice of kernel (similarity function)

linear kernel:

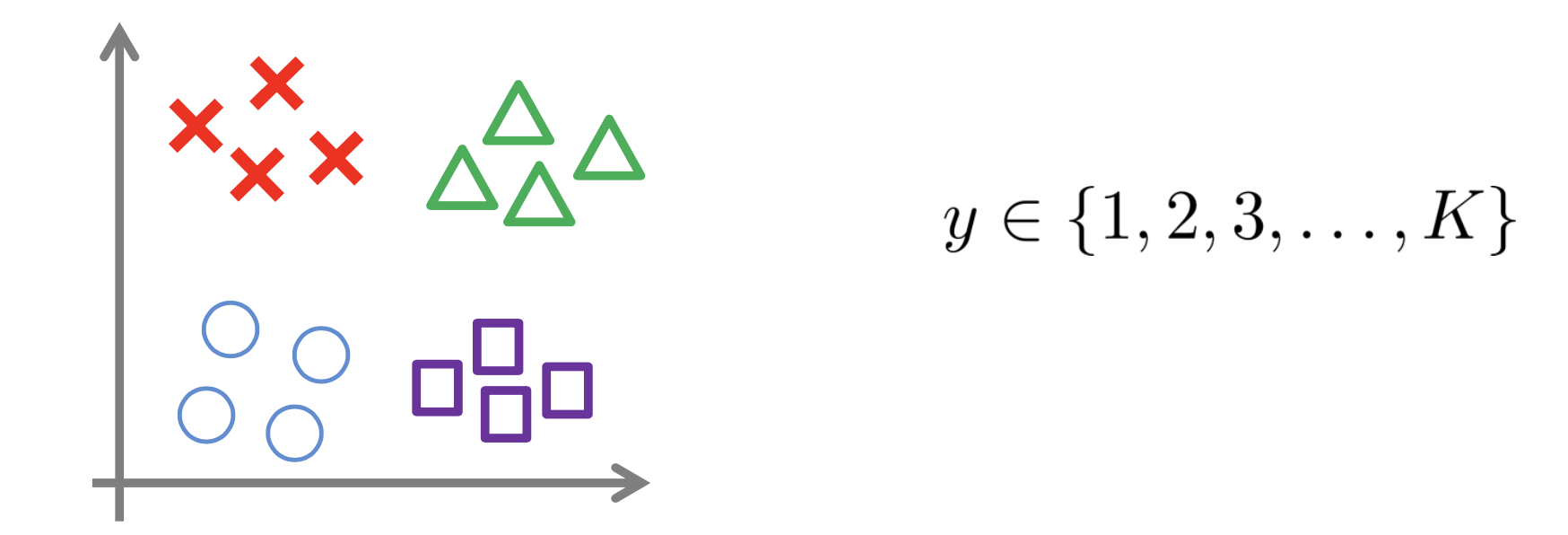


Gaussian kernel:



Need to choose 

**Multi-class classification**



Many SVM packages already have built-in multi-class classification functionality.

Otherwise, use one-vs. -all method. (Train K SVMs, one to distinguish y = i from the rest, for i = 1,2,.....K ), get (1), (2) , …., (K) . Pick class i with largest ((i)) ^ T x

**Logistic Regression vs. SVMs**

N = number of features, m = number of training examples

If n is large (relative to m):

Use logistic regression, or SVM without a kernel (“linear kernel”)

If n is small, m is intermediate:

Use SVM with Gaussian kernel

If n is small, m is intermediate:

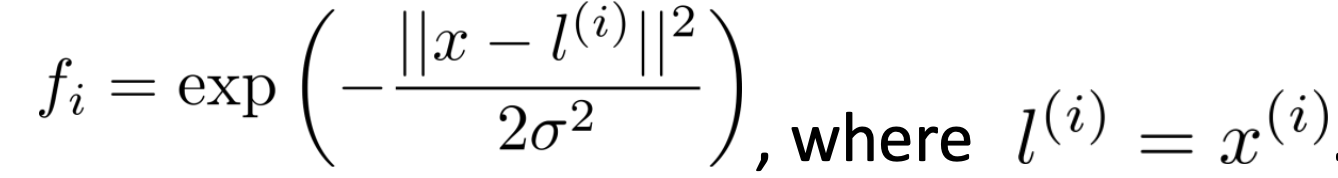
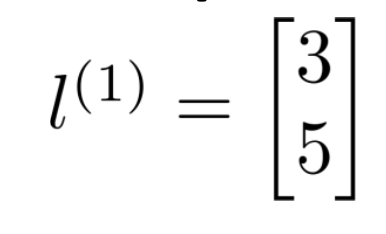
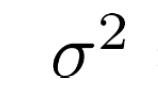
Use SVM with Gaussian kernel

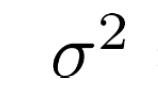
If n is small, m is large:

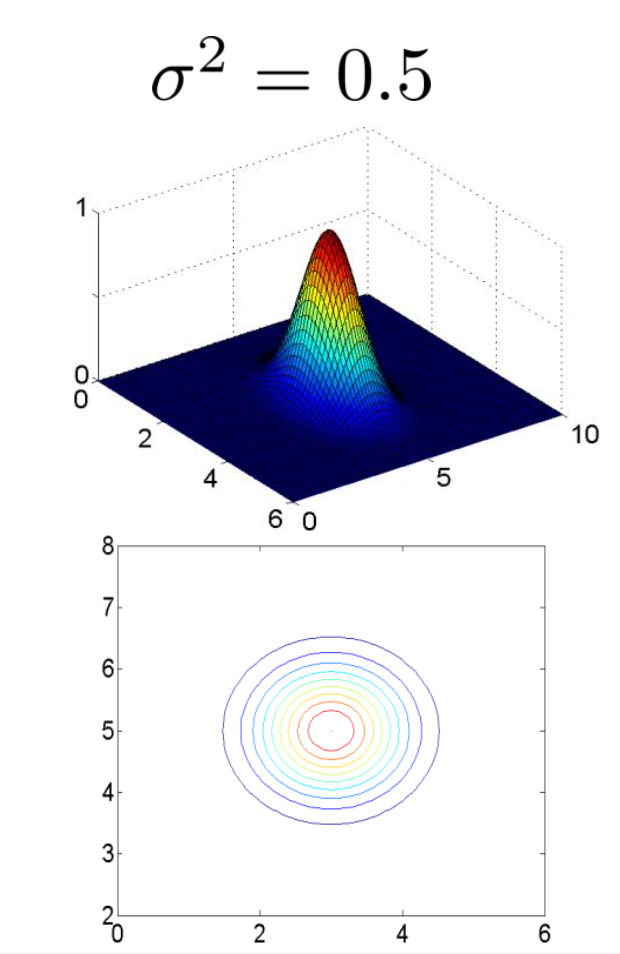
Create/add more features, then use logistic regression or SVM without a kernel.

Neural networks are likely to work well for most of these settings, but may be slower to train.

**Activity:**

What would the 3-D plot of  where i = 1 and , look like for values of  = 1, 0.75 and 5 look like?

An example of such a plot for = 0.5 is as follows:



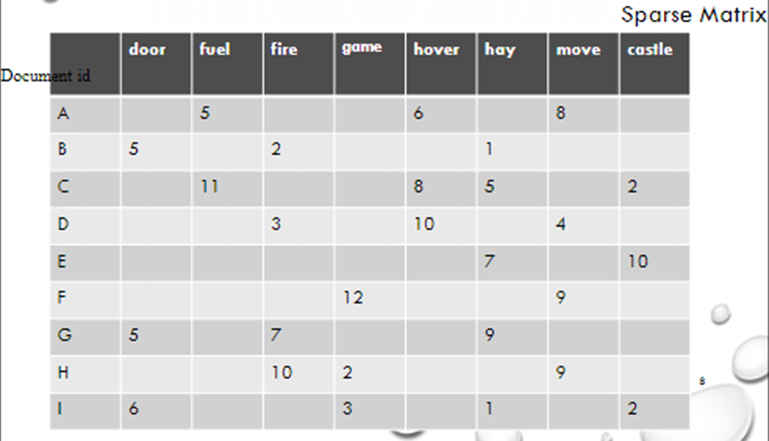
**Reference :**

Machine Learning by Andrew N.G (Coursera)

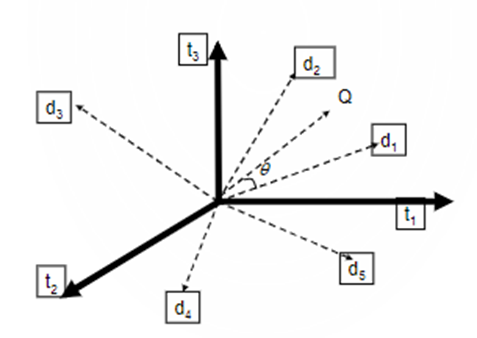
**Vector Space Model (VSM) - By Salman Khan**

VSM refers to the idea and practice of converting documents, and queries, into vectors and obtaining ranked documents, corresponding to an entered query, using their vector representations.

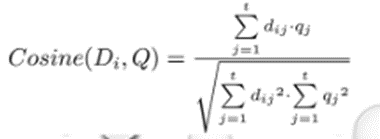
To achieve this, a vector, of the length of words contained in the corresponding index for the corpus, is created for each of the documents contained in the corpus and each word from the index is associated with a location in the vector. Then the vectors are filled with weights for those terms in their corresponding documents, generally using some variation of TF or TF-IDF scheme. The table below (from class slides) shows an example in which each row represents a vector corresponding to the document identifier mentioned in the first column. In the table, the locations in vectors representing words that do not exist in the corresponding documents are empty and can be filled with zeros in vectors.



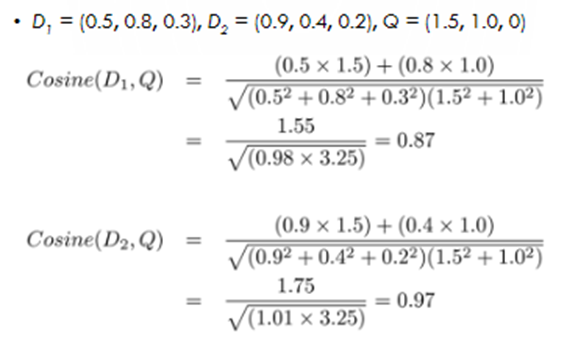
When a query is to be searched for, a vector of the same dimension as the one used for the documents, representing the same terms, is filled for the terms contained in the query. This query vector can then be compared with the vectors representing documents, using cosine similarity, to get a ranked list of the documents based on their similarity with the query. Visually, the representation and comparison of vectors with three terms t1,t2 andt3 can be seen below. In it, the vectors d1, d2, …,di represent the documents and the vector Q represents the query.



Cosine similarity can be computed using the following formula. In it, t represents the total number of documents:



And the following is an example of the calculation:



**VSM Activity**:

DA = (0.6,0.7,0.1); DB = (0.8,0.2, 0.5); DC = (0.9,0.3,0.7)

Q = (0.5,0,0.8)

Based on the vectors representing the documents (A, B and C) and the query (Q) above, rank the three documents using cosine similarity.

# Future Of Information Retrieval

By Edwin Jose and Irene Kahvazadeh

There have long been extensive research possibilities in the area of information retrieval systems, which have been the focus of scholarly attention for a long time. the emergence and evolution of modern information retrieval systems can be divided into six periods [4]:

* To replace original sources, documents were indexed and categorised manually prior to the 1950s;
* During the 1950s, indexing and search changed significantly because of the conditions created after World War II;
* The 1960s saw the development of interactive access and the ARPANET project by the US government during this era of more computer memory;
* Text search began in the 1970s. Dialog, Orbit, Medline, and Lexis are some examples of interactive information retrieval systems developed during this decade;
* In the 1980s, full text, the internet, and search engines were developed; and
* It was during the 1990s that the World Wide Web was born.

Information retrieval and search applications continue to evolve as computing environments change. Social media and mobile devices are the most obvious examples of this type of change in recent years. IR professionals have responded by developing social search, a search method that involves communities of users and informal information exchange. Researchers are developing effective new tools for managing personal and social information through research in areas such as user tagging, conversation retrieval, recommendations, filtering, and collaborative search. Desktop search, which shares many characteristics with the current mobile search applications, was an important early paper in this area [1]. In web IR, short queries with little linguistic structure (typically a single noun compound) have received much attention.

Support has also been provided for users who ask longer, more natural questions. TREC [2] had a question answering task, which dealt with finding simple textual answers to a limited range of questions (such as "who" and "when" questions). It then evolved into the more detailed questions found in large question-answering archives. Additionally, researchers have developed techniques for providing more focused answers to detailed questions. These research findings are partly responsible for Apple's Siri, IBM's Watson, and Yahoo! Answers' success [4].

In the future, certain domains and research in IR could be of great interest:

* Explainable IR especially for Professional searches:

It has been shown that search engine user interaction data can introduce biases, especially related to gender and ethnicity, into IR. Stereotypical biases are reinforced by IR systems when undetected and unidentified users are exposed to them. A second problem is that model retrieval models based on transformer-style over-parameterized models can be fragile and sensitive to small adversarial errors. These statistical over-parameterized models may learn shortcuts due to recent developments in inductive biases, pre-training, and transfer learning practices. Learning patterns that are right for the wrong reasons are generated by shortcuts that are not aligned with human understanding. In addition, expert users using specialized search systems - such as those in legal search, medicine, journalism, and patent search - require control, agency, and lineage. For all the above IR-centric reasons, among many other general reasons – such as utility for legal compliance, scientific investigation, and model debugging – the field of ExIR provides the tools/primitives to examine learning models and the capability to build transparent IR systems [9].

* ChatBots in IR:

Chatbots are computer programs that simulate human communication by using voice commands, text chats, or both. A chatbot is an artificial intelligence (AI) feature that can be embedded and used through any major messaging application [7]. A retrieval-based chatbot relies on predefined responses to a user message in closed-domain scenarios. Intent classification, entity recognition, and response selection constitute the three main functions of a retrieval-based bot [8].

* New Language models and its effect on IR

In the first place, language models are built to predict the likelihood of a pattern or sequence of words occurring. Consequently, NLP uses these models to comprehend language and word predictability. In addition, transfer learning and competent language models in NLP make it easier for machines to learn and understand languages. To predict a sentence's sequence of words, language models utilize techniques in statistics and probability. A language model in NLP evaluates and interprets various text datasets to develop insight for word prediction. Therefore, the languages models offer various features in applications and devices for creating text. Using Language Models, NLP capabilities can be built [6]:

* Machine Translation: It helps translate texts between languages. Google Translate is an example.
* Sentiment Analysis: It provides features to determine emotional responses and behavior. Classifiers for online reviews are an example.
* Text-to-Speech: Furthermore, it can provide voice-based interactive services by transforming text into speech. Alexa is an example.
* Content Categorization: Using the textual data, it builds categories and indexes them in an efficient manner.
* Quantum Inspired Information retrieval

Information retrieval (IR) involves finding relevant information based on an individual's needs. Information consumption and interaction have changed dramatically over the last two decades. Web search engines, easy access to the Internet, and the explosion of information available online have contributed to this change. Various types of information are available - from lecture slides to news articles to descriptions and reviews of items. The IR systems must continually improve to accommodate such growing information needs, both qualitatively (in terms of complexity) and quantitatively. It is possible to reduce IR systems to two basic functions. The first is how to efficiently and effectively represent and rank the variety of unstructured information being created at any given time. Through advanced representation methods, it is possible to index and improve understanding of the content, as well as rank the information items based on the representation. A better understanding of natural language, for example, can improve the representation of textual information. Secondly, IR systems need to be able to better understand users' complex information needs. In order to accomplish this goal, the search context, the search task, and the intent of the user must be understood, as well as the ability to measure the task completion and the satisfaction of the user through user interaction. Research on IR systems has investigated different approaches from both a system and user perspective (representation and ranking). In quantum-inspired IR (QIR), the mathematical framework of Quantum Theory (QT) is utilized to develop representations and user models that are expected to better align with human cognitive information processing. It differs from Quantum Computing in that it does not use quantum states for computations. IR benefits from QT in many ways. By computing probabilities of events, it provides a new way to represent events. In QT, events are represented as subspaces in an abstract, complex vector space (called Hilbert space) rather than as subsets of a set-theoretic sample space. Furthermore, a single event can be represented in more than one Hilbert space basis. As a result of this representation method, information objects like documents and queries can be abstracted and contextualized. When a set of basis vectors corresponds to a set of documents, another set of basis vectors in the same Hilbert space can represent the same set of documents in a different context. Based on the context of retrieval, a query (as an event) will be represented by these different basis. By taking into account interference between events, Hilbert space representations of events lead to a generalized method of calculating probabilities (Born rule). It can be used to model a user's decisions under ambiguity more effectively than traditional probability models. The method inherently models incompatible variables - those where the measurement of one variable affects the outcome of the other.Also, order effect results from incompatibility in measuring human decisions. The application of QT to IR can be broadly divided into two subareas: Representation and Ranking, and User Interaction. A re-ranking task often involves user interactions such as relevance feedback, which overlaps with traditional IR. As QIR also falls under these two sub-areas, it overlaps with traditional IR. QIR uses different tools. In order to model cognitive interference in document ranking, this paper uses the mathematical framework of QT, including complex Hilbert space models for representation learning. IR borrows heavily from concepts, models, and techniques developed in Quantum Cognition, especially in modeling and incorporating user interactions. The QT model has been successfully applied to model and predict irrational human decision making and to explain cognitive biases in human [5].

References:

1. Dumais, Susan, et al. "Stuff I've seen: a system for personal information retrieval and re-use." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. 2003.
2. Voorhees, Ellen M. "The TREC question answering track." *Natural Language Engineering* 7.4 (2001): 361-378.
3. Sanderson, Mark, and W. Bruce Croft. "The history of information retrieval research." *Proceedings of the IEEE* 100.Special Centennial Issue (2012): 1444-1451.
4. Asadnia, Abolfazl, et al. "Future of information retrieval systems and the role of library and information science experts in their development." *Journal of Librarianship and Information Science* (2022): 09610006211067537.
5. Uprety S, Gkoumas D, Song D. A survey of quantum theory inspired approaches to information retrieval. ACM Computing Surveys (CSUR). 2020 Sep 28;53(5):1-39.
6. <https://www.hitechnectar.com/blogs/here-are-the-top-nlp-language-models-that-you-need-to-know/>
7. <https://www.investopedia.com/terms/c/chatbot.asp>
8. <https://www.codecademy.com/learn/retrieval-based-chatbots/modules/retrieval-based-chatbots/cheatsheet>
9. Verberne S. Explainable IR for personalizing professional search. InProfS/KG4IR/Data: Search@ SIGIR 2018 Jan 1.

------------------------------ End Of Future Of Information Retrieval ------------------------------