Analysing Huge Data Collection And Comparing Through Algorithms: KNN, Naive And Collaborative Filtering & Hybrid

Pooja Mudgil, Shivani Gautam, Uditta Chhabra, Mansi Jadaun, Paras Jain, Vikas Singh

Abstract: Recommendation systems are used to obtain and analyse huge datasets of business organisations and industries thus, helping as well as allowing them to identify the best throughput and optimised options for their increase in efficiency and performance. This technology gains its merits in different other technologies for analysis of data. Organisations are able to gain if they are able to recommend suitable products to variant users by use of correct set of tools. Correct product recommended to customers by companies leads to congeniality for either ends. If, used at wide scale can lead to increase in sale of products, increasing profit margins and satisfied customers. This paper presents the effectiveness of recommendation system and its best suitable algorithm that could be used according to the data set available for the corresponding increase in efficiency and productivity by clubbing results from various other researches with the obtained results from analysing of datasets obtained from Kaggle using three algorithms: Naïve Bayes, KNN, and collaborative filtering. For any business, production and growth are in direct correlation with the user's usage and requirements which is successful only when a particular user is able to obtain the products correspondingly at the same time and it could be fast and efficient when the results of recommendation system amplify the user's choices with preferences. Therefore, the studied patterns obtained from researches and through the dataset, implementations of algorithms and comparing them for obtaining an accurate solution for recommendation systems.

Keywords: Recommendation system, Naïve Bayes, K-Nearest Neighbour (KNN), Collaborative filtering, java, Hadoop tool, Hive.

1 INTRODUCTION

Product industry has a vast amount of data which is growing as the demand of different products by the users which is with increase in population increasing at exponential rate day by day. Usage of recommendation system by the industries would turn their traditional way of searching and representing products into a much smarter classification based on certain preferences by the users or users for the products. Thus, allowing huge datasets to be arranged according to the requirements and choices of the users, management as well as its correct usage, helping users as well as industries to overcome the problem of storing and analysing current situation for usage of perfect algorithm. This idea couldn't be possible without uplifting and involvement of databases with its knowledge into the requirements observed by the people. This certainly, now-a-days has been renowned as knowledge discovery in databases. This prior knowledge with machine learning and tools like Hadoop can reduce the working load of brains with a single input and output operation. An open source from Apache that can manage processing, structured and unstructured data as well as storing of application that involve big data providing flexibility towards analysing data and acts as a distributed framework in clustered systems and provides huge support for data mining (through patterns that can generate new information) as in [1], predictive analysation and recommendation purposes as well as machine learning, the newest growing technologies and applications of the world. Visitors over internet who e-commercialise led business to gain its explosive growth. There have been many systems developed based being it Devil Finder and Alta vista that prioritise and personalise according to the information required to be retrieved as in [2].The technique can itself affect cost, revenue, operational efficiency, product manufacturing, liking of most of the users using that same product and preferences, type of product recommended can even give a company to know the high demand and requirement for a particular product and research for upgradation, in all, overall growth and development.

1.1 Recommendation System

Development in technology allows us to move forward towards the easiest and fastest way to achieve and accomplish from the smallest to the convoluted problems. The solution to problems like these has been obtained with the emergence of world-wide-web exploding, the commercial emergence with growth and success leading to development of filtering based technology by determination of N items set in personalised information that might be in interest of a user. This platform or engine that then helps or predicts i.e. recommends through some rating or user preferences based on set of items expressed as recommendation system [3].

1.2 Recommendation system in production industry

The factor which gets affected mostly by success in production industry is the satisfaction of stakeholders involved, thus storing data in their repositories could be used to attain it. Simultaneously if recommendation system gets involved can perfectly suit its purpose, for example: it can verify the most used type or genre of products the customer is using, that further could identify the likeability of finding of the same type of product next time by the same user, then there comes the requirement to use the substance again, products varieties could be involved on the basis of searches and even ratings given by users play an important role in identifying the worth of a substance, a person finds for it. There have been few implementations with lateral growth too in this context but still

- *Pooja Mudgil is currently Assistant Professor of I.T in BPIT (GGSIPU), Delhi, India, E-mail: poojamudgil@bpitindia.com*
- *Shivani Gautam, Uditta Chhabra, Mansi Jadaun, Paras Jain & Vikas Singh are currently pursuing bachelor's degree program in information technology in BPIT(GGSIPU), India,*
  E-mail: shivanigautam19nov@gmail.com

to confine and assess, it requires testing in different circumstances and under various algorithms and problems.

1.3 Contribution of paper

This paper aims to contribute the significance of usage of different algorithms for a data set which turns out to be the best possible outcome as recommendation when implemented to get high efficiency and reduced efforts. In order to show it, a case study (Section 4) has been provided which shows how efficiency of recommendation system can be increased by using the different algorithm at different data sets or situations. In the case study, we have taken an example of book selling site which sites the selling of number of which genre of books, and collection of various books available in the market. In this paper (Section 2) reveals the Literature survey in this area and how different researches has been done over different types of datasets collection for obtaining the varied results allowing the data analysis study much more better with increase in getting the correct product at ease followed by proposed work in (Section 3). A case study has been shown in Section 4 which tells how an algorithm affects the result of recommendations provided by the system when used with KNN, Naïve Bayes, and Collaborative Filtering. Finally we conclude in Section 5 with future scope.

2 LITERATURE REVIEW

Data analytics and manipulation has taken growth with increase in population day by day, increasing the need for variety of products and its development leading the technology and business to take over the analysing of huge data bases for finding out exactly what a customer desires for and how to attract the users using a different product. In [4] simultaneously, shows how usage of various filters with the growth has paced the technology and its efficiency, it shows the user based collaborative filtering based on N set items, analyses for scalability and computes similarity between different items and combines to obtain the similarity between the items and recommender's items. This has been evaluated in the paper with 9 datasets that has obtained the algorithms result as being the proposed item based are up to 2 orders of magnitude faster than the user based neighbourhood systems quality. This again promote to the consideration in variation of collaborative filtering as in [5] for the basis as weighted, mixed, switching and copyright as held to combine different types of recommender systems, provided with an example of classification of movie through IMDB database and obtains that it reaches up to the mark when used with content boosted collaborative filtering. Another research on the algorithms involve about KNN as in [6], k nearest neighbour based algorithm providing the ranking performance such that comparison is done with the weighted distance KNN to decision tree graph and naïve Bayes and advances in with proposing an improvement by combining the KNN with simple naïve Bayes through dealing with the problem of lack of data sets when k is small. Similarly, in [7] it has been shown naïve Bayes being accurate for many of the classification tasks and states that these tasks are for smaller databases and does not scale up as decision tree and simultaneously, paper proposes a new algorithm where hybrid of decision tree and naïve Bayes classifier is considered leaving the comparison analysis with results in scaling up of accuracy when proposed algorithm used.

KNN	Collaborative Filtering	Naive Bayes
Discriminative classifier	Classifies on the basis of other's choices	Eager learning classifier
Supervised as well as lazy classifier, proves difficult to use in prediction for real time	Separates on the basis of similar entities	Assumes conditional independence between features and takes probabilistic estimation for each class
Inherent nature of analysing locally	Content based	Keeps on learning overtime
Usually focused on finding similarity	Data sparsity problem occurs	Much faster than KNN
More complex decision boundaries than complex trees	Limited content analysis is possible	Inherent nature for generative classifier
It over fits with more complexity	Shows scalability problems	Can work best in real time
Accuracy decreases with more complexity	Reduce quality of recommend-dations	Its accuracy increases with more learning
Too slow	Slow start	Faster in comparison to KNN
Can be best with PCA, SVD and usually used in small complexity of $O(n^2)$, n = no, of data points	Similar items based situation are suitable for its usage, usually used by companies like amazon	Can be best in situation where new addition of data keeps on happening for example of email spam with complexity theta(1)

Table 1. *This table compares the three algorithms considered to be suitable for a simple recommender system with variant databases as in [8]*

The paper [9] considered Naïve Bayes algorithm using user interface, helps in prediction and tracking of the pages on the internet based as one of the learning classifier working in a real time and allowing users to help and suggest with the ratings and options. As in [10] the prior success limited popularity of the algorithm KNN and its less usage with the data size increase limits the accuracy. The hybrid algorithm has been used as a solution to KNN and collaborative algorithms. In sparsity and scalability methods, combining the two might give the best accuracy as it can be obtained for the increase in performance and efficiency for a recommender system. In paper [11][12] a technique has been described that compromises between various number of selected variables and performance classifiers. Such that, for measuring the search effectiveness where the need is to find the most relevant output or item and no irrelevant item is retrieved or missed any item that was relevant, this can be done with the help of two basic parameters with in a set considering A = the set of items retrieved and B= the set of relevant items in the database, this first classification considers the intersection as AՈB = relevant items – retrieved ones simultaneously this gives A-B = set of relevant items – not retrieved and simultaneously, B-A = set of all the irrelevant items retrieved – the retrieved ones.

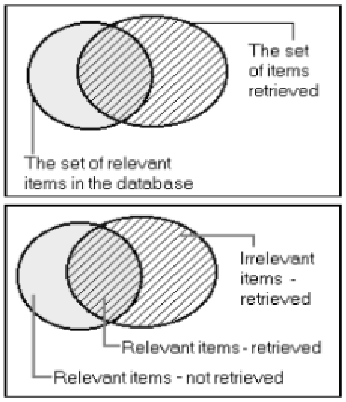

The set of items retrieved
The set of relevant items in the database
Irrelevant items - retrieved
Relevant items - retrieved
Relevant items - not retrieved

Figure 1.*The figure is a representation to the sets we used to classify the data.*

This classification helps in considering our point of measuring the implementations we have done on our recommendation system for books as a case through two variants: Precision and Recall which provides us the accuracy percentage comparison between these algorithms. Where, Recall provides the ratio of the no. of relevant records retrieved to the sum of all the relevant records in the database and expressed in percentage, and is inversely related to the value of precision as in [13] i.e. the ratio of the no. of relevant records retrieved to the sum of all the no. of irrelevant and relevant records retrieved again in percentage.

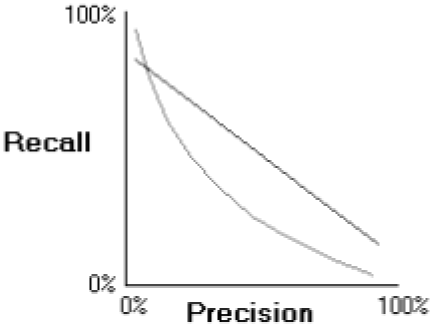


Figure 2. *In the graph above, the two lines represents the performance of different search systems. While the exact slope of the curve may vary between systems, the general inverse relationship between recall and precision remains as in [14].*

Because of this, comprehensive retrieval takes place and this leads to omission of secondary concepts. The problem is complicated when the individual perception is someone may not be relevant to others or any other person. The two aspects are useful measures even they might be having some limitations and thus the figures show the formula used for calculating the values of recall and precision while sets A, B and C are considered for A = no. of relevant records retrieved, B = no. of relevant records not retrieved and C = No. of irrelevant records retrieved.

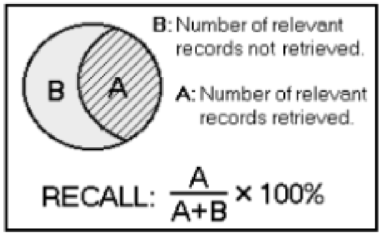

B: Number of relevant records not retrieved.
A: Number of relevant records retrieved.
B
A
RECALL: A / A+B × 100%

Figure 1 *The basic definition used for Recall on the basis of no. of retrieved and relevant records*

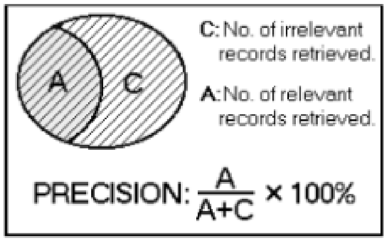
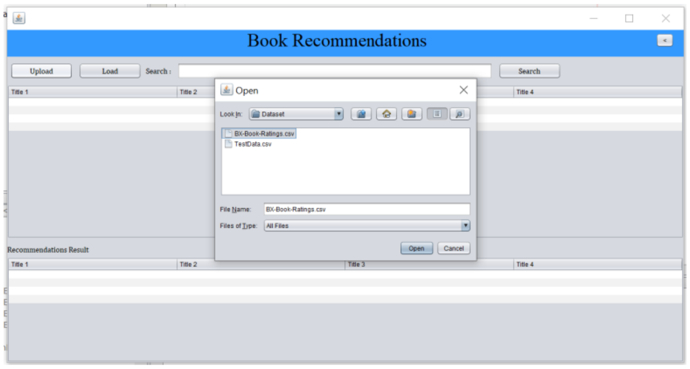

C: No. of irrelevant records retrieved.
A: No. of relevant records retrieved.
A
C
PRECISION: A / A+C × 100%

Figure 2 *The basic definition used for the calculation of precision on the basis of the no. of retrieval of irrelevant and relevant records as in [15].*

These variants can be used and is implemented in the next section showing a case of book recommendation system with the involvement of dataset used as described above. The comparison provides the basis of comparison of implementation and the method of conclusion for this paper.

3 IMPLEMENTATION

After the study of various researches and comparisons, the case of book recommendation system is implemented in this section, with the combination of algorithms to pave way for the best and efficient output obtaining criteria and for the highest accuracy measurement for a system. The below shown screenshots of the outputs obtained and which lead to the conclusion of getting the values of precision and recall we performed for various algorithms as in the following:



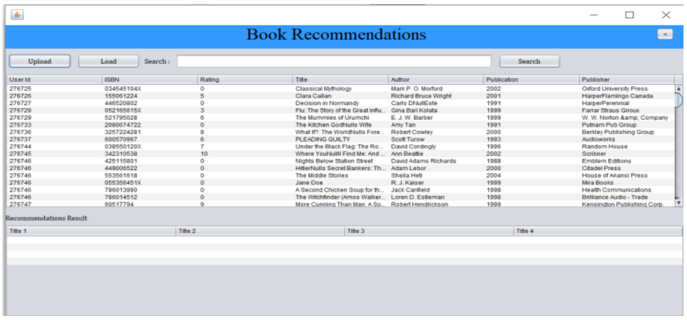
Look In: Dataset

BX-Book-Ratings.csv
TestData.csv

File Name: BX-Book-Ratings.csv

Files of Type: All Files

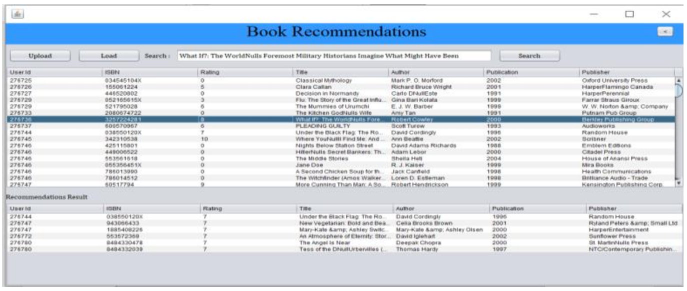
Figure 3. *The above shows the book recommendation system designed on the basis of combination of algorithms and implementing the uploading of datasets into the system.*



User Id	ISBN	Rating	Title	Author	Publication	Publisher
276725	034545104X	0	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press
276726	155061224	5	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada
276727	446520802	0	Decision in Normandy	Carlo DNullEste	1991	HarperPerennial
276729	052165615X	3	Flu: The Story of the Great Influ...	Gina Bari Kolata	1999	Farrar Straus Giroux
276729	521795028	6	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton &amp; Company
276733	2080674722	0	The Kitchen GodNulls Wife	Amy Tan	1991	Putnam Pub Group
276736	3257224281	8	What If?: The WorldNulls Fore...	Robert Cowley	2000	Berkley Publishing Group
276737	600570967	6	PLEADING GUILTY	Scott Turow	1993	Audioworks
276744	038550120X	7	Under the Black Flag: The Ro...	David Cordingly	1996	Random House
276745	342310538	10	Where YouNullll Find Me: And ...	Ann Beattie	2002	Scribner
276746	425115801	0	Nights Below Station Street	David Adams Richards	1988	Emblem Editions
276746	449006522	0	HitlerNulls Secret Bankers: Th...	Adam Lebor	2000	Citadel Press
276746	553561618	0	The Middle Stories	Sheila Heti	2004	House of Anansi Press
276746	055356451X	0	Jane Doe	R. J. Kaiser	1999	Mira Books
276746	786013990	0	A Second Chicken Soup for th...	Jack Canfield	1998	Health Communications
276746	786014512	0	The Witchfinder (Amos Walker...	Loren D. Estleman	1998	Brilliance Audio - Trade
276747	60517794	9	More Cunning Than Man: A So...	Robert Hendrickson	1999	Kensington Publishing Corp.

Title 1	Title 2	Title 3	Title 4

Figure 4. *This shows the dataset uploaded window to be implemented with the algorithms connected in the back end*



Search : What If?: The WorldNulls Foremost Military Historians Imagine What Might Have Been

User Id	ISBN	Rating	Title	Author	Publication	Publisher
276725	034545104X	0	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press
276726	155061224	5	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada
276727	446520802	0	Decision in Normandy	Carlo DNullEste	1991	HarperPerennial
276729	052165615X	3	Flu: The Story of the Great Influ...	Gina Bari Kolata	1999	Farrar Straus Giroux
276729	521795028	6	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton &amp; Company
276733	2080674722	0	The Kitchen GodNulls Wife	Amy Tan	1991	Putnam Pub Group
276736	3257224281	8	What If?: The WorldNulls Fore...	Robert Cowley	2000	Berkley Publishing Group
276737	600570967	6	PLEADING GUILTY	Scott Turow	1993	Audioworks
276744	038550120X	7	Under the Black Flag: The Ro...	David Cordingly	1996	Random House
276745	342310538	10	Where YouNullll Find Me: And ...	Ann Beattie	2002	Scribner
276746	425115801	0	Nights Below Station Street	David Adams Richards	1988	Emblem Editions
276746	449006522	0	HitlerNulls Secret Bankers: Th...	Adam Lebor	2000	Citadel Press
276746	553561618	0	The Middle Stories	Sheila Heti	2004	House of Anansi Press
276746	055356451X	0	Jane Doe	R. J. Kaiser	1999	Mira Books
276746	786013990	0	A Second Chicken Soup for th...	Jack Canfield	1998	Health Communications
276746	786014512	0	The Witchfinder (Amos Walker...	Loren D. Estleman	1998	Brilliance Audio - Trade
276747	60517794	9	More Cunning Than Man: A So...	Robert Hendrickson	1999	Kensington Publishing Corp.

User Id	ISBN	Rating	Title	Author	Publication	Publisher
276744	038550120X	7	Under the Black Flag: The Ro...	David Cordingly	1996	Random House
276747	943066433	7	New Vegetarian: Bold and Bea...	Celia Brooks Brown	2001	Ryland Peters &amp; Small Ltd
276747	1885408226	7	Mary-Kate &amp; Ashley Switc...	Mary-Kate &amp; Ashley Olsen	2000	HarperEntertainment
276772	553572369	7	An Atmosphere of Eternity: Stor...	David Iglehart	2002	Sunflower Press
276780	8484330478	7	The Angel Is Near	Deepak Chopra	2000	St. MartinNulls Press
276780	8484332039	7	Tess of the DNullUrbervilles (...	Thomas Hardy	1997	NTC/Contemporary Publishin...

Figure 5. *The searched results shown as in the figure, After implementation provides the recommended books datasets to the user.*

4 RESULT

These implementations provided with some values which we further implemented as described in literature review for our implementation method concludes with these set of values for different algorithms in the book recommendation system:

	Precision	Recall
Naïve Bayes	0.821	0.081
KNN	0.87	0.851
Collaborative Filtering	0.913	0.092
Hybrid	0.965	0.972

Table 1. *The table defines the values obtained while considering each of the algorithms considering the book recommendation system as a case.*

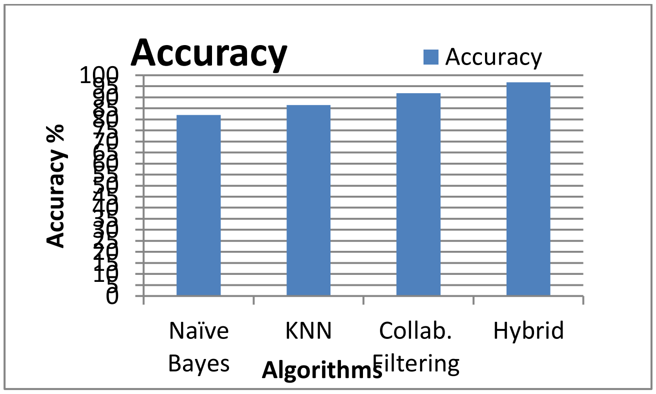


Figure 6 *The above figure gives the chart interpretation of the obtained results on the basis of comparison of Accuracy*

5 CONCLUSIONS

After compiling, book recommendation systems with four different algorithms their accuracy is attained which is mentioned in above figures and results. Hence, it can be concluded that all four algorithms work quite efficiently, where hybrid leads the way with 96% closely followed by collaborative filtering with 92%. It is not that hybrid will always be the best option because there are various factors to be looked upon. Here in this case hybrid works in most efficient manner.

6 FUTURE SCOPES

This conclusion leads to its possibility of reduction in complexity further and can lead us to work for increasing in accuracy methods in future. As well as user interface could be further made much easier and faster through usage of cross platforms like Android or Swift for different substitutes in the market that can be converted considering its implementation and usefulness in daily life can be induced to uplift the technological importance of these type of systems.

REFERENCES

[1]. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases, 17 December, 2008

[2]. P.Resnick, H.R. Varian, Recommender systems published in Commun ACM, 40(3)(1997), pp. 56-58, 10.1145/245108.24512

[3]. L.S. Chen, F.H. Hsu, M.C. Chen, Y.C. Hsu, Developing recommender systems with the consideration of product profitability for sellers published in Int. J Inform Sci, 178(4) (2008), pp. 1032-1048

[4]. Item-Based Top-N Recommendation Algorithms∗ Mukund Deshpande and George Karypis University of Minnesota, Department of Computer Science Minneapolis, MN 55455

[5]. Souvik Debnath, Niloy Ganguly, Pabitra Mitra, “Feature Weighting in Content Based Recommendation System Using Social Network Analysis” WWW 2008, April 21–25, 2008, Beijing, China. ACM 978-1-60558-085-2/08/04.

[6]. Jiang L., Zhang H., Su J. (2005) Learning k-Nearest Neighbor Naïve Bayes for Ranking. In: Li X., Wang S., Dong Z.Y. (eds) Advanced Data Mining and Applications. ADMA 2005. Lecture Notes in Computer Science, vol 3584. Springer, Berlin, Heidelberg

[7]. Ron Kohav, “Scaling up the accuracy of naïve bayes classifiers: A decision tree” KDD’96 Proceedings of the Second International Confernce on knowledge Discovery and Data mining. Pages 202-207

[8]. Pooja Mudgil, Paras Jain, Vikas Singh, “Data Analytics and Data monitoring Based on Database Recommendation – A Comparison”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 1166-1170, March-April 2019. Available at doi: https://doi.org/10.32628/CSEIT1952312, Journal URL: http://ijsrcseit.com/CSEIT1952312.

[9]. P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In 10th national conference on

Artificial Intelligence, pages 223–228. AAAI Press, 1992.

[10]. M.J. Pazzani, A framework for collaborative, content-based and demographic filtering, Artific Intell Rev, 13 (1999), pp. 393-408 No. 5(6)

[11]. The Optimality of Naïve Bayes Harry Zhang Faculty of Computer Science University of New Brunswick Fredericton, New Brunswick, Canada

[12]. J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging: A tutorial. Statistical Science, 14(4):382–417, 1999.

[13]. M. Buckland, F. Gey, The Relationship Between Recall and Precision, Journal of the American Society for Information Science, 45(1):12--19, 1994

[14]. Vijay Raghavan, Peter Bollmann, Gwang S. Jung, " A critical investigation of recall and precision as measures of retrieval system performance", published in Journal ACM Transactions on Information Systems (TOIS) Volume 7 Issue 3, July 1989, pages 205-229

[15]. Bookstein, A.(1974). "The anomalous behaviour of precsion in the Swets model, and its resolution, Journal of Documentation, 21, 374-380.