SISTEM TEMU KEMBALI INFORMASI *(INFORMATION RETRIEVAL SYSTEM)* DOKUMEN BERBAHASA JAWA MENGGUNAKAN METODE *KUMAR-HASSEBROOK (PCE)*

Fatkhul Amin[1], Sariyun Naja Anwar[2], Purwatiningtyas[3]
[1,2,3] Fakultas Teknologi Informasi, Universitas Stikubank Semarang
e-mail: fatkhulamin@edu.unisbank.ac.id, [2] sariyunna@edu.unisbank.ac.id, [3]purwati@edu.unisbank.ac.id

ABSTRAK

Pencarian informasi menggunakan Sistem Temu Kembali Informasi (STKI) atau mesin pencari umumnya menghasilkan hasil pencarian berupa kumpulan dokumen yang besar sehingga membutuhkan waktu untuk menentukan yang paling tepat. Algoritma Kumar-Hassebrook digunakan untuk menghasilkan keluaran (output) berupa dokumen teks bahasa jawa dalam pencarian informasi menggunakan STKI bahasa Jawa dengan didukung stemmer bahasa jawa. Software STKI dirancang untuk memberikan hasil pencarian dokumen dalam jumlah recall rendah dan precision tinggi menggunakan metode pemeringkatan Kumar-Hassebrook, sehingga user akan mendapatkan hasil pencarian cepat dan akurat. Hasil Aplikasi metode Kumar-Hassebrook dengan jumlah data di korpus 400 didapatkan hasil pencarian dengan tingkat rata-rata dokumen terambil (recall) rendah yaitu 0,05 dan tingkat rata-rata akurasi (precision) tinggi yaitu 0,85. Angka akurasi 0,85 menunjukkan bahwa STKI telah bekerja akurat berdasarkan uji recall dan uji precision dengan model persepsi. Tampilan hasil pencarian informasi menempatkan hasil pencarian dengan bobot tertinggi pada posisi paling atas (descending) dengan disertai bobot dokumen dan letak dokumen.

Kata-Kunci: *Kumar-Hassebrook, Bahasa Jawa, STKI*

1. PENDAHULUAN

1.1 Latar Belakang

Bahasa Jawa sebagai bahasa daerah yang paling banyak digunakan di Negara Indonesia, dewasa ini sudah mulai mengalami kemunduran dalam hal jumlah pengguna bahasa jawa itu sendiri. Bahasa jawa tidak lagi menjadi bahasa daerah yang digunakan dalam kehidupan sehari-hari namun sudah mulai tergantikan dengan bahasa Indonesia. Hal ini tentunya sangat mengkhawatirkan dalam rangka pelestarian budaya bahasa jawa khususnya bahasa jawa yang digunakan dalam kehidupan sehari-hari. Secara tidak sadar para pengguna bahasa jawa mengajari anak-anak mereka menggunakan bahasa nasional dalam kehidupannya, jarang sekali menggunakan bahasa jawa. Namun demikian, bukan hanya unsur orang tua saja yang menjadi penyebab berkurangnya pengguna bahasa jawa. Media masa baik yang offline maupun online juga sangat sedikit yang mengangkat bahasa jawa atau menggunakan bahasa jawa dalam medianya.

Beberapa faktor menjadi penyebab bahasa jawa mulai ditinggalkan oleh penggunanya. Ada dua faktor utama yang menyebabkan bahasa Jawa (bahasa daerah pada umumnya) ditinggalkan oleh masyarakat, yaitu faktor internal dan faktor eksternal [1]. Adapun Faktor internal yang dimaksud; Sosialisasi dalam Keluarga yang melemah, Kurikulum Pendidikan yang mengalami disorientasi, dan Kesadaran Generasi Muda yang mulai berkurang. Sedangkan Faktor eksternal yang menjadi penyebabnya yaitu; Modernisasi dan Globalisasi, Eksistensi Bahasa Asing di Indonesia, dan Dominasi Kultural.

Pencarian informasi saat ini dilakukan dengan menggunakan sistem temu kembali informasi (STKI) atau mesin pencari, *user* menuliskan *query* dan mesin pencari akan menampilkanlkan hasil pencarian. STKI yang sudah ada dan banyak digunakan saat ini memberikan hasil perolehan pencarian yang banyak (banyak dokumen yang terambil), sehingga diperlukan waktu untuk menentukan hasil pencarian yang relevan. Menentukan hasil yang relevan sesuai dengan keinginan user dengan jumlah hasil pencarian yang banyak akan menyulitkan pengguna (*user)*. Hal ini terjadi karena dokumen yang terambil oleh sistem jumlahnya banyak, maka sistem berkemungkinan menampilkanlkan hasil pencarian yang tidak relevan. Banyaknya dokumen hasil pencarian ini membuat waktu yang dibutuhkan dalam pencarian menjadi lebih banyak dari yang diharapkan.

Adapun solusi untuk mengatasi masalahini adalah dengan membuat *softwareSistem Temu Kembali Informasi(STKI)* menggunakan MetodeKumar-Hassebrook *yang* didukung oleh *Stemmer* Bahasa Jawa.MetodeKumar-Hassebrook dipilih karena cara kerja model ini efisien, mudah dalam. *SoftwareIRS* diharapkan menghasilkan *recall* rendah dan *precision* tinggi.

2. TINJAUAN PUSTAKA

Penelitian terkait dengan menggunakan metode Kumar-Hassebrook dilakukan oleh Cesar San Martın, dkk,[4] dengan topik "*Improved Infrared Face Identification Performance Using Nonuniformity Correction Techniques*". Metode Kumar-Hassebrookdigunakan untuk mengetahui Tingkat pengenalan wajah menggunakan kinerjaCitra inframerah ditingkatkan dengan menambahkan nonuniformity teknik *preprocessing*. Sistem pencitraan inframerah dapat dibentuk oleh Focal-Plane-Teknologi Array, sekelompok photodetectors terletak di bidang fokusSebuah sistem pencitraan, namun secara inheren menghadirkan ketidaksenonagaman seperti yang

ditetapkan-Pola kebisingan yang menurunkan kualitas gambar inframerah. Selain itu,Keunikan ini perlahan bervariasi dari waktu ke waktu, dan tergantung pada teknologiOgy yang digunakan, drift ini bisa memakan waktu beberapa menit sampai jam. Karena ini, wajahnyaKinerja pengidentifikasian terdegradasi dari waktu ke waktu, membutuhkan continuous-Metode kalibrasi waktu untuk mempertahankan tingkat pengenalan wajahMenggunakan sistem pencitraan inframerah. Dalam sintesis, karya ini berfokus pada evaluasi degradasi dalam kinerja pengenalan pola yang dihasilkanDengan noise pola-fix dan perbaikan ketika nonuniformitas Teknik koreksi diterapkan

Penelitian terkait dengan menggunakan metode *Kumar-hassebrook similarity* dilakukan Sung-Hyuk Cha [9], Survei Komprehensif tentang Jarak/Kesamaan antara Fungsi Probabilitas DensitasUkuran jarak atau kesamaan sangat penting untuk bisa dipecahkan.Banyak masalah pengenalan pola seperti klasifikasi, pengelompokan,Dan masalah pengambilan. Berbagai jarak/kesamaan ukuran ituBerlaku untuk membandingkan dua fungsi kepadatan probabilitas, pdf singkatnya,Ditinjau dan dikategorikan dalam sintaksis dan semantikhubungan. artikel ini membangun bangunan jarak/ kesamaan ukuranDengan menyebutkan dan mengkategorikan berbagai macamJarak/kesamaan ukuran untuk membandingkan tipe nominalHistogram Mengelompokkan tindakan di atasBerkonsentrasi pada tiga aspek umum: kesamaan sintaksis,Implementasi peringatan, dan semantik. metode *Kumar-hassebrook similarity* digunakan menentukan jarak yang sesuai/kesamaan ukuran tidak bisaTerlalu ditekankan Ada permintaan terus-menerus untuk yang lebih baik.

Penelitian tentang DICE Similarity juga dilakukan oleh Khuat Thanh Tung Dkk, [3]mengukur kesamaan dokumenmemainkan peran penting dalam teks terkait penelitian danaplikasi seperti dokumen clustering, deteksi plagiarisme,pencarian informasi, terjemahan mesin dan esai otomatisscoring. Banyak penelitian telah diusulkan untuk memecahkan inimasalah. Mereka dapat dikelompokkan menjadi tiga pendekatan utama:String berbasis, Corpus-based dan Pengetahuan berbasis Persamaan.Dalam tulisan ini, kesamaan dua dokumen yang diukur denganmenggunakan dua ukuran berbasis-string yang berbasis karakter danalgoritma berbasis jangka. Dalam metode berbasis karakter, n-gram adalahdimanfaatkan untuk mencari sidik jari untuk sidik jari dan menampilkanalgoritma, maka koefisien Dice digunakan untuk mencocokkan duasidik jari yang ditemukan. Dalam pengukuran berbasis jangka, cosinusalgoritma kesamaan digunakan. Dalam karya ini, membandingkan efektivitas algoritma yang digunakan untuk mengukurkesamaan antara dua dokumen. Dari hasil yang diperoleh,kita dapat menemukan bahwa kinerj sidik jari dan menampilkanlebih baik dari kesamaan kosinus. Selain itu, menampilkan yangalgoritma lebih stabil daripada yang lain.

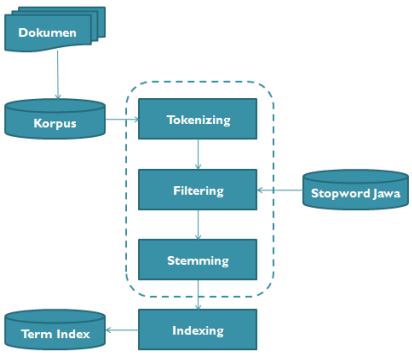
3. METODE PENELITIAN

Penelitian ini menggunakan model *prototype.* Berikut adalah tahapan yang dilakukan pada penelitian ini dengan metode pengembangan *prototype.* Analisa, Pada tahap ini dilakukan analisa tentang masalah penelitian dan menentukan pemecahan masalah yang tepat untuk menyelesaikannya.Menentukan tujuan pembuatan mesin pencari. Disain, Pada tahap ini dibangun rancangan Sistem Temu Kembali Informasi bahasa jawa, *Prototype,* Pada tahap ini dibangun Sistem Temu Kembali Informasi Bahasa Jawa.Tahap ini di mulai dari proses tokenisasi, Penyaringan (filtering), Pembuatan kata dasar bahasa jawa (stemming), tfidf, dan perhitungan *Kumar-Hassebrokk* yang diaplikasikan dengan program PHP. Pengujian, Pada tahap ini dilakukan uji *Recall* dan *Precission* dengan model Persepsi, Evaluasi, Pada tahap ini dilakukan evaluasi apakah performa aplikasi sudah sesuai dengan yang diharapkan, apabila belum maka dilakukan penyesuaian-penyesuaian secukupnya. Penyesuaian Tahap ini dilakukan apabila pada evaluasi performa aplikasi kurang memadai dan dibutuhkan perbaikan, tahap ini melakukan penyesuaian dan perbaikan pada aplikasi sesuia dengan kebutuhan

4. HASIL DAN PEMBAHASAN

4.1. Arsitektur Informasi

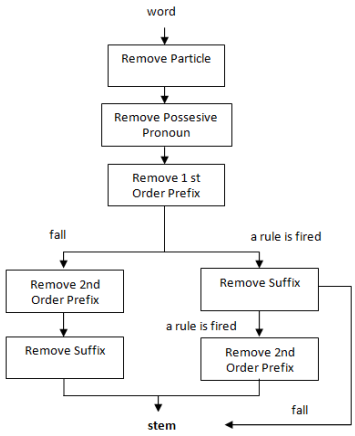
Mesin pencari yang dirancang menggunakan metode Kumar-Hassebrook diharapkan bisa menghasilkan hasil pencarian secara akurat. Arsitektur Informasi pada STKI *Jawa* metode Kumar-Hassebrook Bahasa Jawa menggunakan model Sistem temu kembali Informasi yang hanya diambil pada proses *Pre Processing*. Gambar 1 menunjukkan arsitektur informasi STKI bahasa *Jawa.*



Gambar 1. Arsitektur Informasi Stemmer Bahasa Jawa

4.2. Stemmer Bahasa Jawa

Pada penelitian stemmer bahasa Jawa digunakan arsitektur *stemmer* bahasa Indonesia Tala. Hal ini dilakukan karena Tala melakukan penelitian tentang *stemmer* bahasa Indonesia, dan bahasa jawa adalah ibu dari bahasa Indonesia. Implementasinya menyesuaikan dengan semantik dan tata bahasa yang ada pada bahasa jawa. Gambar 2 menunjukkan *stemmer* bahasa Indonesia



Gambar 2. *The basic design of a Porter stemmerfor*Bahasa Indonesia (Tala, 2003)

4.3. Perancangan Sistem Temu Kembali Informasi (STKI)

STKI dirancang agar mudah digunakan oleh penggunanya (user) dan dirancang dengan tampilan mesin pencari (*Search enginee*) pada umumnya. STKI didesain untuk menemukan informasi secara akurat bagi pengguna [1]. *Flowchart* diawali dengan *input* dokumen-dokumen kedalam korpus. Selanjutnya dokumen melalui proses preprosesing, dihitung bobotnya dan dibuat rankingnya berdasarkan bobot dokumen yang tertinggi. Hasil STKI adalah dokumen yang relevan dengan permintaan *user*.

5. PEMBAHASAN

5.1. Korpus dan preprosessing

Semua dokumen abstrak di *input* secara manual dengan format dokumen teks. Proses ini dilakukan dengan cara memasukkan dokumen teks berbahasa Jawa kedalam tabel korpus. Sebelum dimasukkan kedalam tabel, dibuat satu tabel dengan nama tabel korpus yang digunakan sebagai tempat data. Tabel korpus ini memiliki *fiel-field* id, judul, isi dan dokumen. *Field* id berisi urutan data penelitian didalam korpus yang tersusun sesuai dengan urutan input data. Field judul berisi judul Artikel bahasa jawa. Field isi berisi dokumen jawa dan field dokumen berisi nama dokumen dengan kode tertentu

Proses *scanner* dokumen korpus menggunakan format teks dilakukan dengan cara masuk kedalam dokumen korpus melalui perantara program php ke dalam database mysql. Proses *scanner* data dilakukan dengan cara *scanner* baris per baris, untuk tiap-tiap file naskah yang ada di dokumen. Tokenisasi dimulai dengan memisahkan *term-term* yang ada pada dokumen korpus menjadi kumpulan term melalui proses *scanner* dengan dasar spasi.

Proses selanjutnya setelah proses tokenisasi adalah proses *filtering*. Proses filtering dilakukan untuk menghilangkan *term-term* yang tidak memiliki arti dengan menggunakan *stopword list* tala.Proses filtering adalah proses baca tabel kedua untuk diperiksa apakah semua term memiliki term-term yang termasuk dalam stopword list menurut tala. Jika dalam tabel kedua terdapat *term-term* yang termasuk dalam *stopword*, maka akan dilakukan penghilangan *term-term* tersebut.Proses menghilangkan kata atau term yang tidak memiliki makna dilakukan dengan cara membuang *stopword (stopword removal)* dilakukan untuk menghilangkan *term-term* yang tidak memiliki arti dengan menggunakan *stopword jawa* .

5.2. Indexing

Proses *indexing* dilakukan untuk mengambil atau *retrievedterm-term* yang ada pada tabelfreq untuk selanjutnya diproses pada saat pencarian dilakukan oleh STKI. Proses perhitungan dilakukan langsung pada STKI saat *query* diproses oleh sistem. Pengguna (*User*) memasukkan Kata Kunci (*query*) pada mesin pencari, kemudian setelah kata kunci ditulis mesin pencari akan melakukan pencarian *query* pada *database* dengan mengolahnya terlebih dahulu sesuai dengan arsitektur mesin pencari menggunakan metode *Kumar-Hassebrook* dan memberikan hasil pencarian.

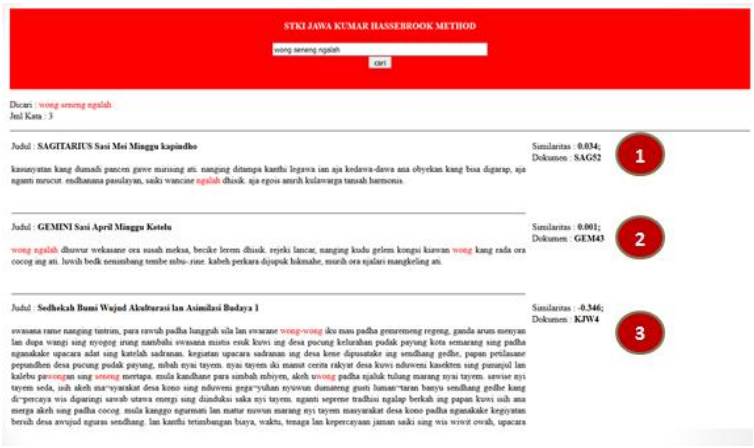
5.3. Hitung Similaritas Metode Kumar-Hassebrook

Kumar-Hassebrookmerupakan metode yangdigunakan untuk menghitug tingkat kesamaan(similarity) antar dua buah objek. Untuknotasi himpunan dapat digunakan rumus (1):

$$s_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} \quad (1)$$

dimana p dan q adalah dokumen yang berbeda. p_i adalah term i yang ada di dokurnen p q_i adalah term i yang ada di dokumen q

Studi kasus pada aplikasi STKI ini menggunakan dokumen-dokumen teks sejumlah 300 dokumen Palintangan Basa Jawa pada Majalah Online Penjebar Semangad yang terdapat pada 12 Palintangan (zodiak) yaitu; Capricorn, Aquarius, Pisces, Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio, dan Sagitarius. *Query* yang dimasukkan pada STKI adalah *keyword* dengan 2 *term* yaitu "*seneng ngalah*" , "seneng sabar" 3 *term* "*wong angel sukses*". " wong seneng ngalah" 4 *term* "*wong sukses seneng ngalah*". 5 *term* "*wong sing sukses seneng ngalah*". Gambar 3 menunjukkan hasil pencarian untuk keyword "wong seneng ngalah".



Gambar 3. Hasil Pencarian

Hasil pencarian dokumen dengan keyword "wong *seneng ngalah*", menunjukkan dokumen dengan bobot tertinggi adalah dokumen letak dokumen SAG52 (bobot 0,034). Dokumen SAG52 (dokumen Palintangan Sagitarius Sasi Mei Minggu Kapindo nomer 52) memiliki bobot tertinggi atau memiliki tingkat kemiripan tertinggi dibandingkan dengan dokumen lain yang ada pada korpus.

5.4. Uji Recall dan Precision

Pendeklarasian penting dilakukan untuk proses perhitungan uji recall dan precision. Semua keyword yang akan di uji harus dideklarasikan dulu makna atau arti dari keywordnya. Hal ini penting agar proses pengkategorian kata kunci relevan atau tidak relevan menjadi benar dan ada dasarnya. Adapun dasarnya yaitu dari pencocokan makn yang ada di kamus besar bahasa indonesia (KKBI). Deklarasi persepsi bisa dilihat pada tabel 1

Tabel 1. Tabel Persepsi

Jawa	Indonesia	DEKLARASI (sumber KKBI, https://kbbi.web.id)
sabar	Sabar	sabar/sa-bar/ a 1 **tahan menghadapi cobaan** (tidak lekas marah, tidak lekas putus asa, tidak lekas patah hati); tabah: ia menerima nasibnya dengan --; hidup ini dihadapinya dengan --; 2 **tenang; tidak tergesa-gesa; tidak terburu nafsu:** segala usahanya dijalankannya dengan --;
becik	Baik	baik /ba·ik / 1 a **elok; patut; teratur (apik, rapi, tidak ada celanya, dan sebagainya)**
Seneng	Bahagia	bahagia/ba-ha-gia/ 1 n **keadaan atau perasaan senang dan tenteram (bebas dari segala yang menyusahkan):** -- dunia akhirat; hidup penuh --; 2 a **beruntung; berbahagia**
Ngalah	Mengalah	mengalah/me-nga-lah/ v **mengaku kalah; dengan sengaja kalah (menyerah); tidak mempertahankan pendapat (tuntutan dan sebagainya);**
Wong	Orang	orang n 1 manusia (dalam arti khusus); 2 manusia (ganti diri ketiga yang tidak tentu): jangan lekas percaya pada mulut --; 3 dirinya sendiri; manusianya sendiri: saya tidak bertemu dengan -- nya; 4 kata penggolong untuk manusia: lima -- nelayan; 5 anak buah (bawahan): mereka itu -- nya Pak Camat; 6 rakyat (dari suatu negara); warga negara: -- Pakistan; 7 manusia yang berasal dari atau tinggal di suatu daerah (desa, kota, negara, dan sebagainya): dia -- Bogor; suaminya -- Eropa; 8 suku bangsa; 9 manusia lain; bukan diri sendiri; bukan kaum (golongan, kerabat) sendiri: jangankan anak sendiri, anak -- pun saya tolong; negeri -- , negeri lain (bukan negeri kita); 10 cak karena (sebenarnya):
Angel	Sulit	sulit/su-lit/ a 1 sukar sekali; susah (diselesaikan, dikerjakan, dan sebagainya): pekerjaan yang -- diselesaikan; rasanya -- baginya untuk memberitahukan hal itu kepadamu; 2 susah dicari; jarang terdapat: obat semacam itu -- didapat; 3 dirahasiakan (sukar diketahui dan sebagainya); tersembunyi: tempat -- pun ia tahu; ia dapat mengetahui hal yang --; 4 gelap (rahasia, tidak terang-terangan): apa yang mereka lakukan itu merupakan perbuatan yang --; 5 dalam keadaan yang sukar (genting, gawat, dan sebagainya):
Sukses	Sukses	sukses/suk·ses/ /suksés/ a berhasil; beruntung:

Pengujian *recall(P)* dan *precision (R)* dilakukan dengan cara *inputquery* ke dalam STKI *input* 1 *term*, 2 *term* dan 3 *term, 4 term, dan 5 term*. Perhitungan *recall* dan *precision* menggunakan persamaan (2) dan persamaan (3). Hasil pengujian *recall* dan *precision* dengan menguji 1 *term*, 2 *term* dan 3 *term* sampai dengan 7 term menunjukkan bahwa jika *recall* rendah maka *precision* akan tinggi, selengkapnya terlihat pada tabel 1.

Hasil perhitungan Recall untuk keyword "wong seneng ngalah" adalah sebagai berikut;

$$Recall\ (R) = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ relevant\ items\ in\ collection} \quad (2)$$

$$Recall = \frac{2}{100} = 0.02$$

Hasil perhitungan Precision untuk keyword "seneng ngalah" adalah sebagai berikut;

$$Precision\ (P) = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved} \quad (3)$$

$$Precision = \frac{2}{3} = 0.67$$

Hasil perhitungan rata-rata untuk Recall dan precision adalah sebagai berikut;

$$Rata - rata\ Recall = \frac{0.36}{7} = 0.05$$

$$Rata - rata\ Precision = \frac{6.82}{7} = 0.85$$

5.5. Hasil Uji Recall dan Precision

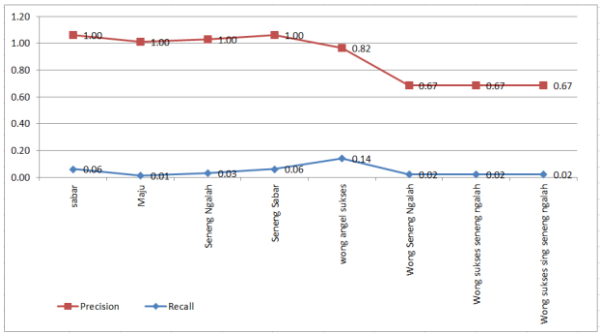
Semua term atau kata diuji menggunakan formula recall dan precision. Hasil uji ditunjukkan pada tabel 2.

Tabel 2 Hasil Pengujian *Recall* dan *Precision*

No	Query	Recall	Precision
1	sabar	0.06	1.00
2	Maju	0.01	1.00
3	Seneng Ngalah	0.03	1.00
4	Seneng Sabar	0.06	1.00
5	wong angel sukses	0.14	0.82
6	Wong Seneng Ngalah	0.02	0.67
7	Wong sukses seneng ngalah	0.02	0.67
8	Wong sukses sing seneng ngalah	0.02	0.67

5.6. Diagram Uji Recall dan Precision

Hasil uji recall dan precision selengkapnya ada pada gambar 4.



Gambar 4. Diagram Hasil Uji

5. KESIMPULAN

Dokumen teks bahasa jawa dengan jumlah data di korpus sebesar 400 dokumen telah dilakukan perhitungan menggunakan metode Kumar-Hassebrook. Berdasarkan analisa persepsi dengan cara dideklarasikan, Hasil Uji *recall* dan *precision* STKI Jawa Metode Kumar-Hassebrook menunjukkan hasil pencarian dokumen teks memiliki rata-rata *recall* = 0,05 dan rata-rata *precision* = 0,85. Artinya STKI yang telah dibangun menunjukkan bahwa STKI akurat. STKIjawa yang dibangun memiliki keunggulan mampu melakukan pencarian dokumen teks bahasa jawa dan hasil pencarian yang akurat (*precision* = 0,85), serta dilengkapi dengan bobot dan letak dokumen pada database

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Stikubank Semarang yang telah memberi "**dukungan financial**" terhadap penelitian ini.

DAFTAR PUSTAKA

[1] Amin, Fatkhul, dkk. 2017. A Hybrid Method Of Rule-Based And String Matching Stemmer For Javanese Language. Journal of Theoretical and Applied Information Technology. ISSN: 1992-8645 . 15th October 2017 -- Vol. 95. No. 19 -- 2017

[2] Kadir, A., 2001. Dasar Pemrograman Web Dinamis menggunakan PHP. Penerbit Andi. Yogyakarta

[3] Khuat Thanh Tung, dkk. 2015.A Comparison of Algorithms used to measure the Similarity between two documents. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 4, April 2015
[4] Martin, CS. 2008. Improved Infrared Face Identification Performance Using Nonuniformity Correction Techniques. ACIVS 2008, LNCS 5259, pp. 1115–1123, 2008. Springer-Verlag Berlin Heidelberg 2008
[5] Manning, C., Raghavan, P., 2007. An Introduction to Information Retrieval, Stanford. USA
[6] Meadow, C.T., 1997. *Text Information Retrieval Systems*. Academic Press.New York.
[7] Tala, F.Z., 2003*, A Study of Stemming Effects on Information Retrieval in bahasa Indonesia*. Institut for logic, Language and Computation Universiteit van Amsterdam The Netherlands.
[8] Salton, G., 1989, *Automatic Text Processing, The Transformation, Analysis, and Retrieval of information by computer*. Addison – Wesly Publishing Company, Inc. USA.
[9] Sung-Hyuk Cha, 2007. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, International Journal of Mathematical models and Methods in Applied Science, issue 4 volume 1.
[10] Yates, R.B, 1999. *Modern Information Retrieval*, *Addison Wesley*-Pearson international edition, Boston. USA