

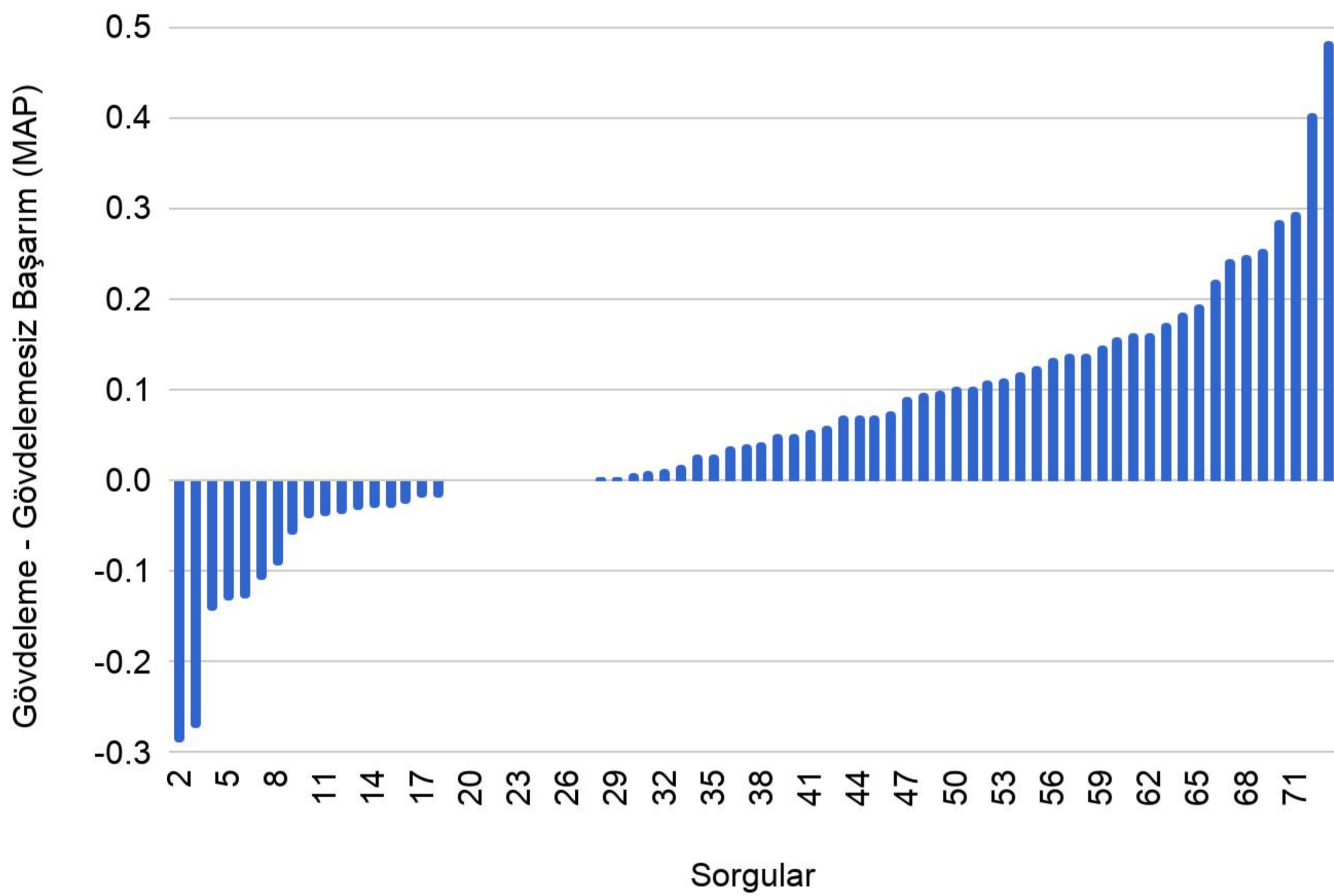
ÖZET

Gövdeleme bilgi erişiminde kullanılan standart tekniklerden biridir. Ancak bu alanda yapılmış çalışmalarda, gövdelemenin bazı sorgularda başarımı artırdığı bazı sorgularda ise düşürdüğü raporlanmıştır. Gövdeleme, sistemin tüm sorgu kümesi üzerinde ortalama başarımda az bir değişim göstermektedir. Seçkili bilgi erişim sistemleri yaklaşımları her bir sorgu için farklı bir bilgi erişim tekniğinin uygulanmasını temel almaktadır. Önerilen seçkili gövdeleme yaklaşımı, koleksiyondaki sözcüklerin gövdeleme uygulanarak ve orijinal halleri ile üretilen iki ayrı indeksten sorguya göre seçim yapan bir mekanizmadır.

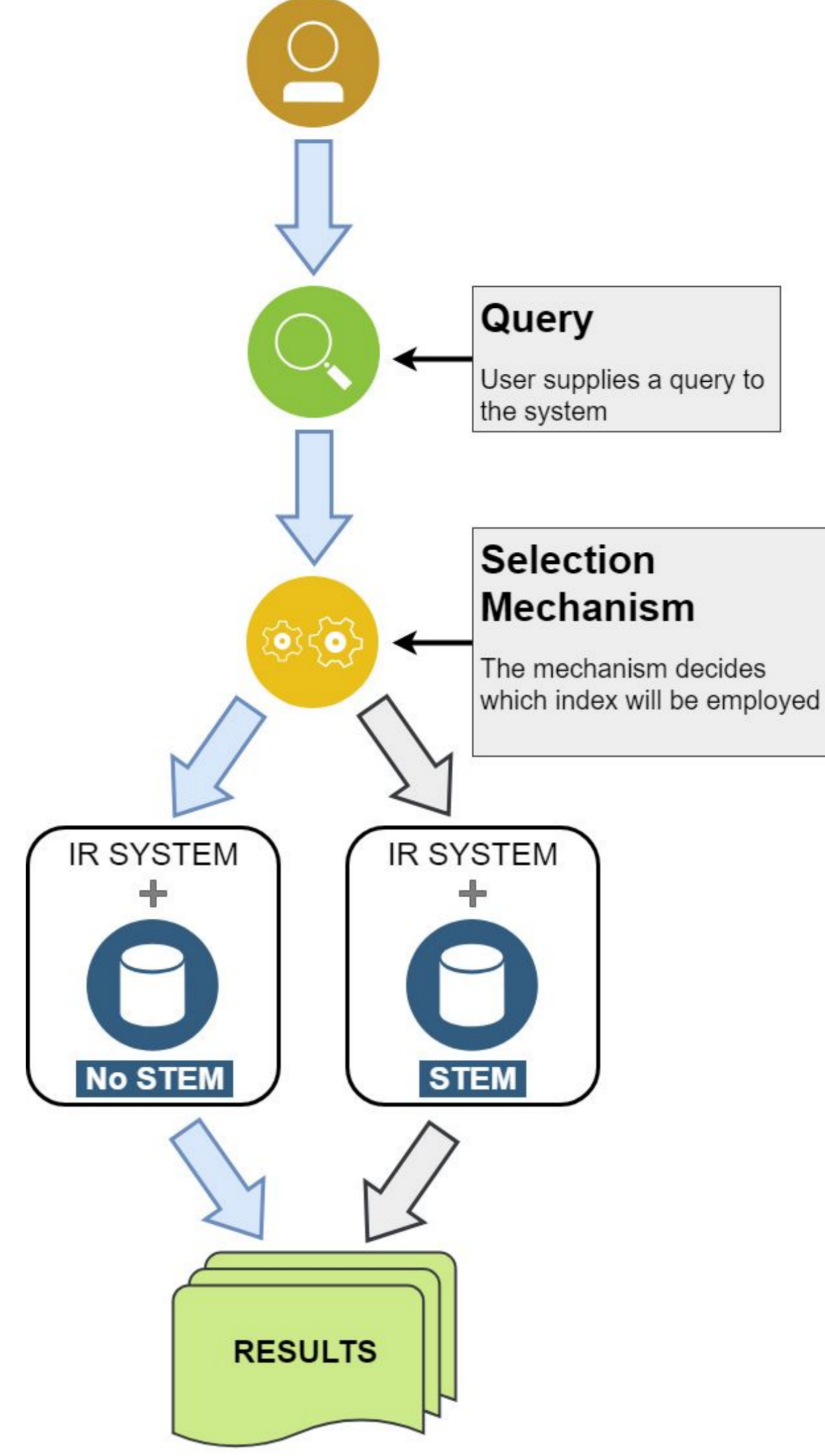
Problem

Gövdeleme Doğal Dil İşlemede bir normalizasyon tekniğidir ve bilgi erişim sistemlerinde de sıklıkla kullanılmaktadır. Gövdeleme ile "Geliyorum", "Geleceğim", "Gelirim" gibi sözcükler "Gel" sözcüğüne indirgenir. Bilgi erişim sistemlerinde sözcüklerin gövdenmesi ile sözcükler orijinal hallerinden genel bir kavrama dönüştürülür. Bu durumda kullanıcının yaptığı sorguya uygun bir çok doküman getirilirken sorguya uygun olmayan alakasız dokümanlarda gövdeleme sebebi ile getirilir. Genel olarak anma (recall) yükselirken duyarlılık (precision) düşer. Ancak gövdeleme bazı sorgularda sistemin genel başarımını artırırken bazı sorgularda ise düşürmektedir [1,2]. Şekil 1, Milliyet Koleksiyonu [3] üzerinde F5 gövdeleme uygulandığında her bir sorgu için gerçekleşen performans değişimini göstermektedir. Bazı sorgularda performans artış gözlenirken bazı sorular için performans düşüşü gözlenmiştir.

F5 Gövdelemenin Milliyet Koleksiyonunda Sorgu Bazında Etkisi



Şekil 1. F5 Gövdelemenin Milliyet Koleksiyonunda Sorgu Bazında Etkisi

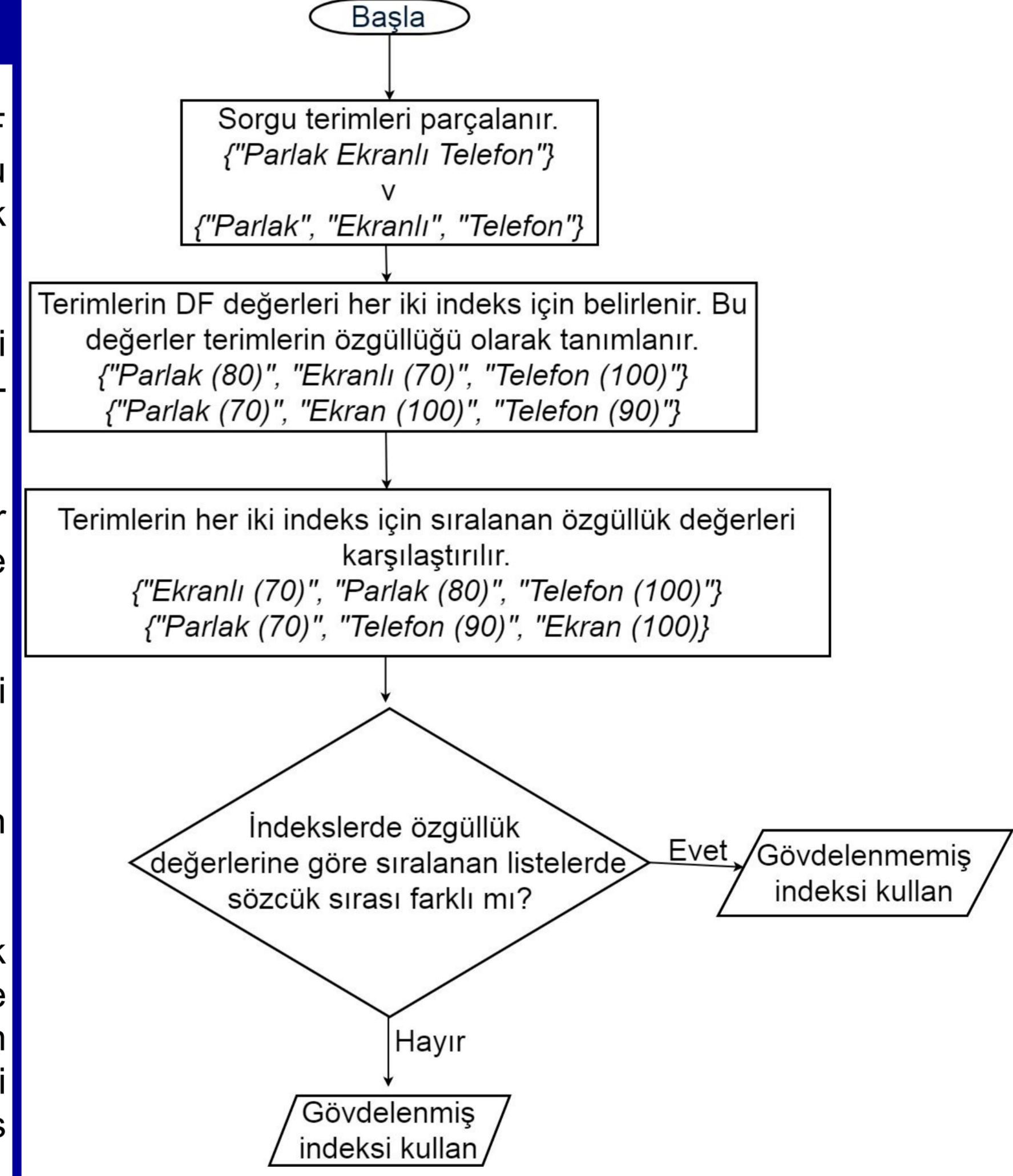


Bilgi Erişim Sisteminin Yapısı

- Sunulan seçkili bilgi erişim sisteminde dokümanlar iki ayrı indekslenmiş veriler bulunmaktadır. Bunlardan biri indeksleme aşamasında uygulanan gövdeleme diğeri ise gövdeleme uygulamadan yapılan indekslemedir.
- Kullanıcının sisteme verdiği sorgu için sorgu terimlerinin özgüllük sıralamalarındaki değişime bağlı olarak hangi indeksin kullanılacağına karar verilir.
- Özgüllük sıralaması iki ayrı indeks için yapılır.
- Özgüllük sıralaması sorgu terimlerinin doküman ve koleksiyon frekanslarına göre yapılır.
- İki ayrı indeks ile oluşturulan sıralamanın karşılaştırılması ile hangi indeksin kullanılacağına karar verilir.

Seçim Mekanizması

- Sorgu terimlerinin DF veya TF değerleri her iki indeks için bulunur. Bu değerler terimlerin özgüllük sıralamasında kullanılır.
- Frekans ve terimlerin sorgu içindeki pozisyonlarından oluşan, anahtar - değer ikilileri üretilir.
- İki ayrı indeks ile oluşturulan bu ikililer anahtara göre sıralanır ve iki ayrı liste oluşturulur.
- Listelerdeki her bir girdinin değerleri karşılıklı olarak karşılaştırılır.
- Bu karşılaştırmaya göre hangi indeksin kullanılacağına karar verilir.
- Bu iki sıralama birbirinden farklılık gösteriyorsa gövdeleme uygulanmayan indeks üzerinde işlem yapılır. Değişimin gözlemlenmediği durumda gövdeleme yapılan indeks üzerinde işlem yapılır.



ClueWeb09A ve NDCG@20 başarımları ile DF sıralamaları

ClueWeb09A	NDCG@20	DF Sıralaması
DirichletLmc200.0(197)	KStem(0.13692)	SelectiveStemming(0.13679)
DFIC(197)	KStem(0.16997)	SelectiveStemming(0.16886)
DLH13(197)	SelectiveStemming(0.11159)	KStem(0.10950)
PL2c2.0(197)	SelectiveStemming(0.14877)	KStem(0.14834)
DPH(197)	KStem(0.15580)	SelectiveStemming(0.15545)
BM25k1.4b0.5(197)	NoStem(0.14982)	SelectiveStemming(0.14536)
DFRee(197)	SelectiveStemming(0.13305)	KStem(0.13258)

Modeller için sorguların başarımları yüksek olan indekslerdeki miktarları

Modeller için	DF Sıralaması	DF Sıralaması
DirichletLmc200.0(0.14441)	NoStem(140)	KStem(57)
DFIC(0.17971)	NoStem(139)	KStem(58)
DLH13(0.11677)	NoStem(158)	KStem(39)
PL2c2.0(0.15664)	NoStem(143)	KStem(54)
DPH(0.16671)	NoStem(136)	KStem(61)
BM25k1.4b0.5(0.16026)	NoStem(149)	KStem(48)
DFRee(0.14157)	NoStem(141)	KStem(56)

KAYNAKÇA

- [1] Harman, D. (1991). How effective is suffixing?. Journal of the american society for information science, 42(1), 7.
- [2] Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2004, November). A framework for selective query expansion. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (pp. 236-237). ACM.
- [3] Can F., Kocberber, S., Balcik E., Kaynak C., Ocalan H. C., Vursavas O. M. First large-scale information retrieval experiments on Turkish texts. In SIGIR 2006, 627-628, 2006.