Doküman Sınıflandırmada Anahtar Kelime Tabanlı Sosyal Ağ Benzerliğinin Deneysel Analizi
Experimental Analysis of Keyword-based Social Network Similarity Approach for Document Classification

Furkan Goz, Osman Kabasakal, Alev Mutlu
Kocaeli Üniversitesi
Bilgisayar Mühendisliği Bölümü
Kocaeli, Türkiye
furkan.goz@kocaeli.edu.tr, okabasakal88@gmail.com, alev.mutlu@kocaeli.edu.tr

***Özetçe* —Çevrimiçi dokümanların artması ile birlikte dokümanlar arası benzerliklerin hesaplanması bir araştırma problemi haline gelmiştir. Bu çalışmada anahtar kelimelere dayalı Sosyal ağ benzerlik yönteminin doküman sınıflandırmasına uygulanabilirliği incelenmiştir. Farklı özellikteki iki veri kümesi üzerinde yapılan deneylerde anahtar kelime tabanlı benzerlik hesaplaması yaklaşımının iyi sonuçlar verdiği görülmüştür.**

Anahtar Kelimeler*—*Doküman benzerliği, PositionRank, Sosyal ağ tabanlı benzerlik.

***Abstract*—With increasing amount of online documents, similarity calculation for such documents has became a challenge. In this study, we focus on document classification and experimentally analyse applicability of keyword based social network similarity for this purpose. Experimental results based on the two different datasets show that applicability of the method is promising.**

Keywords*—*Document similarity, PositionRank, Social network-based similarity.

I. Giriş

Bilgi teknolojisinin gelişmesiyle birlikte çevrimiçi dokümanların sayısı gün geçtikçe artmaktadır. Bu artış dokümanların sınıflandırılması, dokümanların kümelenmesi, dokümanlardan bilgi çıkarımı, doküman özetleme ve doküman önerme gibi birçok metin madenciliği probleminin hızlı ve yüksek başarımla çözülmesi ihtiyacını doğurmuştur [1]. Bu problemlerin başarımını etkileyen temel faktörler doküman temsillerinin oluşturulması ve bu temsillere göre de dokümanlar arası benzerliğin hesaplanmasıdır [2].

Literatürde dokümanlar arası benzerlik hesaplama çalışmaları iki ana sınıfta incelenebilir. İlk sınıfa düşen çalışmalar, dokümanlar arası benzerlik için anlam bilimsel yaklaşımları kullanmaktadır [3], [4]. İkinci sınıfa düşen çalışmalar ise istatistikler yöntemler kullanarak dokümanlar arası benzerlikleri hesaplamaktadır [5]. İstatistiksel yöntemlerin kullanıldığı çalışmaların bir kısmında dokümanı oluşturan tüm kelimeler göz önünde bulundurulurken [5]–[7] bir kısmında da dokümanlardan çıkarılan anahtar kelimeler kullanılmaktadır [6], [7]. Benzerlik hesaplamalarında ise kosinüs benzerliği, Jaccard katsayısı, Öklid uzaklığı gibi yöntemler sıklıkla kullanılmıştır [2], [8].

Bu çalışmada anahtar kelime tabanlı doküman sınıflandırma problemine odaklanılmıştır. Bu amaçla, haber zinciri oluşturmada kullanılan sosyal ağ benzerliği yönteminin [9] anahtar kelime tabanlı doküman sınıflandırma problemine uygulanabilirliği incelenmiştir. Sosyal ağ benzerliği haber zincirlerinin oluşturulmasında varlık isimlerini kullanan Dice katsayısı tabanlı bir benzerlik hesaplama yöntemidir. Bu çalışmada, varlık isimleri yerine dokümanları temsil eden anahtar kelimeleri kullanılmış ve yöntemin başarısı incelenmiştir. Dokümanları temsil eden anahtar kelimeler PositionRank [10] algoritması kullanılarak elde edilmiştir.

Anahtar kelime tabanlı sosyal ağ benzerlik yönteminin doküman sınıflandırmadaki başarısını incelemek için iki veri kümesi üzerinde deneyler yapılmıştır. İlk veri kümesi HAIS konferansında 2017, 2018 ve 2019 yıllarında kabul edilen çalışmaların özetlerinden oluşmaktadır. Bu veri kümesinin özelliği çalışmaların genellikle benzer araştırma problemleri hakkında olmasıdır. İkinci veri kümesi ise birbirinden oldukça farklı konuların işlendiği 5 farklı akademik konferansa ait çalışmaların özetlerinden oluşmaktadır. Her iki veri kümesi üzerinde ikişer deney yapılmıştır. İlk deneyde çalışmaların özetlerini oluşturan tüm kelimeler kullanılarak dokümanlar sınıflandırılmış, ikinci deneyde ise sınıflama işlemi dokümanların anahtar kelimeleri kullanılarak yapılmıştır. Deneysel sonuçlar, anahtar kelimeye dayalı doküman sınıflandırmanın dokümana ait tüm kelimeleri kullanılarak yapılana göre daha başarılı olduğunu göstermektedir.

Çalışmanın devamı şu şekilde organize edilmiştir: II. bölümde önerilen yöntemi oluşturan temel kavramlar anlatılmaktadır. III. bölümde önerilen yöntem ayrıntılı olarak sunulmak-

Bu çalışma TÜBİTAK tarafından 117E566 nolu proje kapsamında desteklenmiştir

978-1-7281-7206-4/20/$31.00

tadır. Deneysel sonuçlar IV. bölümde verilmektedir. Sonuçlar V. bölümde yer almaktadır.

II. Temel Kavramlar

Bu bölümde çalışmada kullanılan PositionRank algoritması ve Sosyal ağ benzerlik yöntemi anlatılmaktadır.

A. PositionRank

PositionRank, ağırlıklandırılmış çizge tabanlı, gözetimsiz bir anahtar kelime bulma algoritmasıdır. PositionRank algoritması, PageRank algoritmasını kelimelerin doküman içindeki pozisyonlarını da dikkate alacak şekilde modifiye etmektedir. Algoritma üç aşamadan oluşmaktadır: (a) dokümanı temsil eden çizgenin oluşturulması, (b) düğüm ağırlıklarının hesaplanması ve (c) anahtar kelimelerin çıkarılması.

Dokümanlar için dokümanı oluşturan her tekil kelime çizgede bir düğüme denk gelecek şekilde çizge oluşturulur. Dokümandaki kelimeler *w* uzunluğundaki pencereler halinde ele alınır ve aynı pencere içine düşen kelimeleri temsil eden düğümler arasına kenar konur. Kenarların ağırlıkları başlangıçta 0 olarak atanır ve kenarların ağırlıkları bağladıkları kelimelerin beraber geçtikleri her pencere için 1 artırılır.

Düğümlerin ağırlıkları ise Denklem 1'e göre hesaplanır. Denklemde, α düğümler arası geçişi sağlayan sabit bir ağırlığı temsil etmektedir. $\tilde{p_i}$, v_i düğümünün doküman içindeki pozisyonlarından dolayı sahip olduğu ağırlığı göstermektedir. w_{ji}, v_j ve v_i düğümleri arasındaki kenarın ağırlığını, $O(v_j)$ ise v_j düğümünde bağlı tüm kenarların ağırlıklarının toplamını vermektedir. Düğüm ağırlık hesabı iki iterasyon arasında düğümlerin ağırlık farkı belli bir değerin altına düşene kadar ya da iterasyon sayısı önceden belirlenmiş bir değere ulaşana kadar devam eder. Başlangıçta tüm düğümlerin ağırlıkları $\frac{1}{|V|}$, $|V|$ çizgedeki düğüm sayısı olmak üzere, atanır.

$$S(v_i) = (1-\alpha)\cdot \tilde{p_i} + \alpha \cdot \sum_{v_j \in Adj(V_i)} \frac{w_{ij}}{O(v_j)} S(v_j) \quad (1)$$

Çizgedeki en yüksek ağırlığa sahip n kelime dokümanın anahtar kelimesi olarak belirlenir. Eğer bu n kelimeden herhangi ikisi ya da üçü metin içinde art arda geliyorsa bu kelimeler birleştirilerek anahtar kelime öbekleri oluşturulur.

B. Sosyal ağ tabanlı benzerlik

Sosyal ağ tabanlı dokümanlar arası benzerlik yaklaşımı sosyal ağlardaki aktörler arasındaki ilişkinin hesaplanmasına dayanmaktadır. Çizge yaklaşımlı bir yöntem olan sosyal ağ benzerlik ölçütünde, dokümanlardaki varlık isimleri birer düğüm olarak kabul edilir. Düğümleri bağlayan kenarların ağırlıkları ise Denklem 2'de verilen Dice katsayısına göre hesaplanır. Denklem 2'de w_{ij} v_i ve v_j düğümlerini bağlayan kenarın ağırlığını, N_c *i* ve *j* varlık isimlerinin beraber geçtiği dokümanların sayısını, N_i ve N_j de sırasıyla *i* ve *j* varlık isimlerinin geçtiği dokümanların sayısını verir.

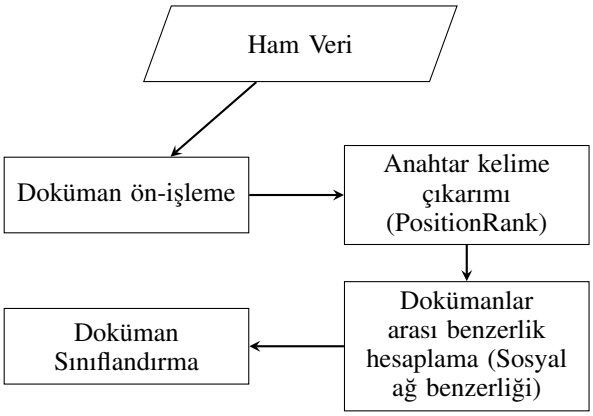
$$w(a,b) = \frac{2 \times N_c}{N_a + N_b} \quad (2)$$

İki dokümanın benzerliği ise, Denklem 3'de gösterildiği gibi, bu dokümanlarda geçen tüm ikili varlık isimlerinin benzerliklerinin toplamının dokümanlarda geçen toplam varlık isim sayısının ikili kombinasyonuna oranı şeklinde hesaplanır.

$$sim(d_i, d_j) = \frac{\sum_{a \in d_i} \sum_{b \in d_j} w(a,b)}{N_p} \quad (3)$$

III. Yöntem

Önerilen yöntem dört ana adımdan oluşmaktadır. Figür 1'de de gösterildiği gibi bu adımlar, (i) veri ön-işleme, (ii) anahtar kelime çıkarımı, (iii) dokümanlar arası benzerlik değerinin hesaplanması ve (iv) doküman sınıflandırma. Bu adımlar aşağıdaki alt başlıklarda açıklanmıştır.



Şekil 1: Sistemin Mimarisi

A. Doküman ön-işleme

Literatürdeki çalışmaların çoğunda dokümanı temsil eden anahtar kelimeler bulunurken sadece dokümandaki sıfatlardan ve isimlerden yararlanılmıştır [11], [12]. Bu çalışmada da veri ön-işleme olarak Stanford NLP kütüphanesi[1] kullanılarak dokümanlardaki kelimeler analiz edilmiş, sıfat ve isim dışındaki kelimeler dokümanlardan silinmiştir.

B. Anahtar kelime çıkarımı

Bu aşamada, veri ön-işleme aşamasında elde edilen yeni dokümanlar PositionRank algoritmasına girdi olarak verilmiştir. Her doküman için, dokümandaki tüm kelimelerin ağırlıkları hesaplanmıştır. Bir dokümandaki anahtar kelimeleri bulmak için öncelikle o dokümanı oluşturan tüm kelimelerin ağırlıklarının ortalaması hesaplanmış ve bu ortalamanın üstünde ağırlık değerine sahip kelimeler o doküman için anahtar kelime olarak kabul edilmiştir.

C. Doküman arası benzerlik hesaplaması

Bu aşamada sosyal ağ benzerlik yöntemi kullanılarak iki doküman arası benzerlik dokümanlar için elde edilen anahtar kelimeler kullanılarak hesaplanmıştır.

Sosyal ağ benzerlik hesaplamasında d_i ve d_j dokümanlarının benzerliği hesaplanırken d_i ve d_j'de yer alan varlık

[1] https://nlp.stanford.edu/software/

isimlerinin ikili benzerlik değerleri diğer dokümanlar da dikkate alınarak bir benzerlik değeri hesaplanmaktadır. Ancak bu yaklaşım aşağıdaki probleme neden olabilir. v_a ve v_b, d_i ve d_j dokümanlarında birlikte sıklıkla geçen ancak diğer az sayıda dokümanlarda birlikte ve birçok dokümanda ayrı ayrı geçen iki varlık ismi olsun. Bu durumda v_a ve v_b kelimelerinin benzerlik değerleri düşük ve d_i ve d_j dokümanlarının benzerliklerine katkısı sınırlı olacaktır. Önerilen yöntemde ise her doküman için elde edilen anahtar kelimeler sadece o dokümanı oluşturan kelimeler dikkate alınarak bulunmaktadır. Bu da yukarıda belirtilen problemin oluşmasını engellemektedir.

D. Doküman sınıflandırma

Bu aşamada sınıfı bilinmeyen bir dokümanın sınıfı bilinen dokümanlarla olan ikili benzerliği hesaplanır. Dokümanın sınıfa olan benzerliği de o sınıfa ait olan dokümanlara benzerlik toplamanının o doküman sınıfındaki doküman sayısına bölünmesi ile elde edilir. Sınıfı bilinmeyen doküman en yüksek benzerlik değerine sahip olan sınıfa atanır. Sınıf belirleme işlemi Denklem 4 ve 5'te gösterilmiştir.

$$sim(d, G_i) = \frac{\sum_{d_i \in G_i} sim(d, d_i)}{|G_i|} \tag{4}$$

$$C(d, G) = \underset{G_i \in G}{\operatorname{argmax}}(sim(d, G_i)) \tag{5}$$

IV. Deneysel Sonuçlar

Bu bölümde önerilen yöntemin başarımını ölçmek için kullanılan veri kümelerinin tanımı ve deneysel çalışmanın sonuçları sunulmaktadır.

A. Veri Kümeleri ve Deney Düzeneği

Önerilen yöntemin başarımını ölçmek için akademik konferanslarda yayımlanan bildirilerin özetlerinden iki veri kümesi oluşturulmuştur. İlk veri kümesi 2017, 2018 ve 2019 yıllarında HAIS[2] konferansında (International Conference on Hybrid Artificial Intelligence Systems) yayımlanan bildirilerin özetlerinden oluşmaktadır. Bu veri kümesinde veri madenciliği ve bilgi keşfi (VMBK), biyo-ilham modeller (BIM), öğrenme algoritmaları (ÖA), görselleştirme (GRS), veri madenciliği uygulamaları (VMU) ve hibrit uygulamalar (HU) isimli oturumlar altında sunulan bildirilerin özetleri yer almaktadır. Oturum isimlerinden de anlaşılacağı gibi bu veri kümesindeki dokümanlar işledikleri konu ve içerik açısından oldukça benzerdir.

İkinci veri kümesi için bilgisayar mühendisliğinin farklı araştırma problemlerine yönelik yapılan konferansların özetlerinden oluşmaktadır. Bu veri kümesindeki dokümanlar 5GWN[3] - kablosuz ağlar, IPTA[4] - görüntü işleme, RoboSoft[5] - robotik, DMBD[6] - veri madenciliği ve SPACE[7] - güvenlik alanlarına yöneliktir ve işledikleri konu ve içerik bakımından da oldukça farklıdır.

[2]https://dblp.org/db/conf/hais/index
[3]https://dblp.org/db/conf/5gwn/
[4]https://dblp.org/db/conf/ipta/
[5]https://dblp.org/db/conf/robosoft/
[6]https://dblp.org/db/conf/dmbd/
[7]https://dblp.org/db/conf/space/

Tablo I'de veri kümelerindeki kategori isimleri, özet bildiri sayısı ve özetlerde bulunan ortalama kelime sayısı bilgileri verilmektedir.

TABLO I: Veri Kümelerinin Özellikleri

	Kategori	#Doküman	#Kelime
V. Kümesi 1	VMBK	28	150
	BIM	28	143
	ÖA	30	120
	GRS	23	164
	VMU	37	150
	HU	41	143
V. Kümesi 2	5GWN	63	127
	IPTA	81	143
	RoboSoft	113	170
	DMBD	73	142
	SPACE	42	171

Tablo II'de veri ön-işleme sonrasında oluşan dokümanlardaki ortalama kelime sayısı ve doküman başına elde edilen ortalama anahtar kelime sayıları verilmektedir.

TABLO II: Veri Ön-işleme Sonuçları

	Kelime sayısı	
	Veri Kümesi 1	Veri Kümesi 2
Ortalama kelime sayısı	97,69	51,28
Ortlalama anahtar kelime sayısı	64,49	33,66

Anahtar kelimelerin doküman sınıflandırmadaki başarımı incelemek için veri kümeleri üzerinde iki farklı deney gerçekleştirilmiştir. İlk deneyde özetin ön-işleme sonucu elde edilen kelimeler kullanılmıştır. İkinci deneyde ise özet için elde edilen anahtar kelimeler kullanılmıştır.

Yöntemin başarısını test etmek için birini dışarıda bırakmaya dayalı çapraz doğrulama yöntemi kullanılmış ve her veri kümesinin başarısı ayrı ayrı değerlendirilmiştir. Bunun için her veri kümesine ait olan her bir özet sırasıyla kümeden çıkarılmıştır. Bu özetin veri kümesindeki sınıfına olan benzerlik değerleri hesaplanmış ve en yüksek değere sahip olduğu sınıf ile etiketlenmiştir.

B. Deneysel Sonuçlar

Tablo III ve Tablo IV'de deneylere ait karmaşıklık matrisleri, Tablo V'te de deneylerin doğruluk sonuçları verilmiştir. Tablo V'te Tüm K. sütunu altında veri ön-işleme sonucu elde edilen kelimeler ile yapılan deneylerin, Anahtar K. başlığı altında ise anahtar kelimeler kullanılarak yapılan deneylerin doğruluk sonuçları verilmiştir.

Veri kümesi 1 için Tablo V incelendiğinde 6 kategoriden 4'ünde anahtar kelime tabanlı yaklaşımın doğruluk oranını artırdığı görülmektedir. Doğruluk oranı özellikle tüm kelimeler tabanlı yaklaşımın kötü sonuç verdiği kategoriler için iyileştiği görülmektedir. Ancak tüm kelime tabanlı yaklaşımın başarılı olduğu öğrenme algoritmaları kategorisinde ise düşüş olduğu görülmektedir. Ancak Tablo III'te verilen karmaşıklık matrisi incelendiğinde iki yönteme de dayalı deneylerde birçok çalışmanın öğrenme algoritmaları kategorisi ile ilişkilendirildiği görülmektedir. Bunun temel nedeni tüm kategorilerdeki özetlere ait anahtar kelimelerin öğrenme algoritmaları kategorisindeki özetlere ait anahtar kelimelerle büyük oranda uyuşmasıdır.

TABLO III: VERİ KÜMESİ 1 KARMAŞIKLIK MATRİSİ

(a) Tüm kelimeler

	BIM	VMU	VMBK	HU	ÖA	GRS
BIM	11	0	1	2	14	0
VMU	6	3	8	3	16	1
VMBK	1	2	8	0	16	1
HU	9	3	7	1	21	0
ÖA	2	0	5	0	23	0
GRS	1	1	4	2	15	0

(b) Anahtar kelimeler

	BIM	VMU	VMBK	HU	ÖA	GRS
BIM	14	1	2	2	8	1
VMU	4	7	9	7	8	2
VMBK	4	6	6	1	10	1
HU	16	7	1	4	11	2
ÖA	3	2	7	3	15	0
GRS	2	4	4	4	5	4

TABLO IV: VERİ KÜMESİ 2 KARMAŞIKLIK MATRİSİ

(a) Tüm kelimeler

	5GWN	IPTA	RoboSoft	DNBD	SPACE
5GWN	53	5	0	4	1
IPTA	0	79	0	0	2
RoboSoft	0	1	112	0	0
DMBD	4	26	2	41	0
SPACE	4	2	1	2	33

(b) Anahtar kelimeler

	5GWN	IPTA	RoboSoft	DNBD	SPACE
5GWN	56	2	0	4	1
IPTA	3	67	3	4	4
RoboSoft	0	0	112	1	0
DMBD	8	7	1	55	2
SPACE	3	0	1	1	37

TABLO V: DOĞRULUK SONUÇLARI

	Kategori	**Tüm K.**	**Anahtar K.**
V. Kümesi 1	VMBK	29	21
	BIM	39	50
	ÖA	77	50
	GRS	0	17
	VMU	8	19
	HU	3	10
V. Kümesi 2	5GWN	84	84
	IPTA	98	83
	RoboSoft	99	99
	DMBD	56	75
	SPACE	78	88

Veri kümesi 2'ye ait 5 kategoriden 2'si için doğruluk sonuçları anahtar kelime tabanlı yaklaşım kullanıldığında artmış, 2'si için değişmemiş ve 1'i için de azalmıştır. Veri kümesi 1 ile karşılaştırıldığında bu deneyin başarım oranları çok daha yüksektir. Bunun da temel nedeni özetlere ait anahtar kelimelerin birbirinden oldukça farklı olmasıdır.

Sonuçlar incelendiğinde anahtar kelime tabanlı sosyal ağ benzerlik yönteminin doküman sınıflandırmada tüm kelimeleri dikkate alan sosyal ağ yöntemine göre 11 deneyden sadece 3'ünde daha kötü sonuç vermiştir, diğerlerindeyse sonuçlar ya aynı ya da daha iyi çıkmıştır.

V. SONUÇLAR

Bu çalışmada anahtar kelime tabanlı sosyal ağ benzerlik yönteminin doküman sınıflandırmadaki başarısı deneysel olarak incelenmiştir. Deneysel sonuçlar anahtar kelimelere dayalı sosyal ağ yönteminin doküman sınıflandırmada kullanılabileceğini göstermektedir.

Gelecek çalışmalarda, yöntemin akademik çalışma dışındaki dokümanlar için uygulanabilirliği incelenecektir. Ayrıca farklı doküman benzerlik hesaplama yöntemlerinin anahtar kelimelerle elde edeceği doğruluk oranları incelenecektir.

KAYNAKLAR

[1] V. Gupta, G. S. Lehal *et al.*, "A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60–76, 2009.

[2] M. D. Lee, B. Pincombe, and M. Welsh, "An empirical evaluation of models of text document similarity," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 27, no. 27, 2005.

[3] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016, pp. 165–174.

[4] S. Wang and R. Koopman, "Clustering articles based on semantic similarity," *Scientometrics*, vol. 111, no. 2, pp. 1017–1031, 2017.

[5] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: Tf-idf approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE, 2016, pp. 61–66.

[6] F. Ricca, E. Pianta, P. Tonella, and C. Girardi, "Improving web site understanding with keyword-based clustering," *Journal of Software Maintenance and Evolution: Research and Practice*, vol. 20, no. 1, pp. 1–29, 2008.

[7] S. Ramasamy and K. Nirmala, "Disease prediction in data mining using association rule mining and keyword based clustering algorithms," *International Journal of Computers and Applications*, vol. 42, no. 1, pp. 1–8, 2020.

[8] R. Mihalcea, C. Corley, C. Strapparava *et al.*, "Corpus-based and knowledge-based measures of text semantic similarity," in *Aaai*, vol. 6, no. 2006, 2006, pp. 775–780.

[9] C. Toraman and F. Can, "Discovering story chains: A framework based on zigzagged search and news actors," *Journal of the Association for Information Science and Technology*, vol. 68, no. 12, pp. 2795–2808, 2017.

[10] C. Florescu and C. Caragea, "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1105–1115.

[11] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[12] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text mining: applications and theory*, vol. 1, pp. 1–20, 2010.